s

# Exploring the bias in very large automatic speech recognition systems

*Simon Knight*


4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2025

# Abstract

This project attempts to detect and analyse bias in the results of a Very Large ASR, specifically Whisper by OpenAI. This testing is completed by comparing WER values of transcriptions generated by Whisper on data from the Speech Accent Archive as well as from the Edinburgh International Accent corpus. The results of L2 speakers are classified using Language family and genus groups. Multiple tests are conducted on multiple Whisper model sizes that extensively explore factors that illuminate obscure vectors through which Whisper could demonstrate Bias against certain linguistic backgrounds.

The results were analysed and groups that Whisper appeared to demonstrate significantly worse performance on were highlighted, with discussion about the potential causes of the bias, or patterns expressed between groups also explored.The results were also statistically validated to see if the distributions of WER values were significantly different.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Simon Knight*)

# Acknowledgements

I would like to thank my family and my siblings, Jonathan, Robert, Andrew and Eleanor, for being an inspiration to me these past 4 years. They make me who I am more than anything else. I would also like to thank friends in Edinburgh and elsewhere, Your support has been critical. I would also like to thank my ever-suffering Supervisor Peter Bell, who has advised me through the course of this dissertation in spite of my habitual tardiness.

# Table of Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 Problem Statement and Contribution . . . 1
- 1.3 Dissertation Structure . . . 2

**2 Background** **3**
- 2.1 Large Automatic Speech Recognition Systems . . . 3
  - 2.1.1 Whisper . . . 3
- 2.2 Bias in ASR . . . 6
- 2.3 Related Work in Bias in ASR . . . 7
  - 2.3.1 Explorations on the cause of ASR Bias . . . 7
  - 2.3.2 Selection of descriptive work on ASR bias . . . 8

**3 Data** **11**
- 3.1 Datasets Selection . . . 11
- 3.2 Edinburgh International Accents of English Corpus . . . 12
  - 3.2.1 Preprocessing . . . 12
- 3.3 GMU Speech Accent Archive . . . 13
  - 3.3.1 Retrieval and Preprocessing . . . 13

**4 Methodology** **14**
- 4.1 Whisper . . . 14
  - 4.1.1 Models . . . 14
  - 4.1.2 Language in the Whisper Model . . . 15
- 4.2 Languages . . . 16
  - 4.2.1 Noise Generation and Mixing . . . 17
- 4.3 Normalisation . . . 18
- 4.4 Calculation . . . 19
- 4.5 Statistical Tools . . . 20
- 4.6 Details of Experimental Process . . . 20
  - 4.6.1 SAA Analysis . . . 21
  - 4.6.2 EdAcc Analysis . . . 21

**5 Results and analysis** **22**
- 5.1 Baseline Analysis . . . 22
  - 5.1.1 results on the SAA . . . 22

5.1.2 Analysis of Base transcripts on the EdAcc corpus . . . . . 28
5.2 Analysis of difference in transcription quality between different versions of Whisper . . . . . 31
5.2.1 Testing on the SAA . . . . . 31
5.2.2 Testing on the EdAcc . . . . . 33
5.3 Tests on the Whisper model's language detection . . . . . 34
5.3.1 Testing on the SAA . . . . . 34
5.3.2 Testing on the EdAcc . . . . . 36
5.4 Whisper on audio samples with noise . . . . . 36
5.4.1 Testing on the SAA . . . . . 36
5.4.2 Testing on the EdAcc . . . . . 37

**6 Conclusions** **39**

**Bibliography** **41**

**A First appendix** **45**

# Chapter 1

# Introduction

This research is concerned with the Exploration and analysis of bias in very large Automatic Speech Recognition (ASR) systems, specifically with OpenAI's Whisper model

## 1.1 Motivation

Automatic Speech Recognition Technology (ASR) has seen remarkable advancements in recent years, with newer models making transcriptions faster, and with fewer inaccuracies. They are finding use in a variety of applications including automated human-machine interactions in services such as call centers, domestic voice assistants, Search Engines, and automated subtitling of videos for the hearing impaired. The newest models use larger amounts of training data than ever before to do this, but it is important to consider the effects that the unintended behaviours of the models may have. This Paper intends to examine bias in ASR systems especially with emphasis on OpenAI's Whisper, a widely adopted system, and aims to build on the body of work exploring fairness and ethics in ASR technology.

ASR technology's primary aim is to provide efficient and accurate transcriptions across all demographics. If there is great variability in the performance of an ASR system based on dialect or accent, it will lead to dramatic variance in user experience, which raises concerns about fairness and bias.

At the inception of this dissertation, OpenAI's Whisper, one such advanced ASR system was approaching 1 year since release and represented one of the best widely available ASR systems. Whisper achieved significant improvements in the accuracy of speech recognition under many circumstances, but still is not immune from biases that can affect its performance for groups based on dialect or accent.

## 1.2 Problem Statement and Contribution

This Research seeks to understand and quantify biases in Whisper, aiming to add to the wealth of literature analysing ASR systems. More specifically We aim to achieve the 4

research goals below:

- Present a broad analysis of bias in Whisper
- Investigate in detail Whisper bias against L2(second language English) speakers
- Investigate bias demonstrated in different sizes of the Whisper model
- Investigate Bias caused by Whisper's langauge detection Functionality
- Investigate Bias in Whisper for noisy audio

These goals were developed organically during the research stage. For example, the focus on L1(first language English) and L2 speakers is due to the SAA corpus' design being condusive to that test, whereas reading OpenAI's Whisper papers benchmarking tests helped in deciding to test bias in noisy environments

In this report I have successfully acquired test datasets for the Investigation, Completed Comprehensive test on the 'base' model of Whisper, and categorised L2 speakers into wider lingusitic groups in order to make more statistically valuable observations. I have also completed tests on Whisper model sizes up to Medium, as well as completing tests on the Whisper's language testing functionality. Finally, I have also completed tests on Whisper at a range of Noise levels, ranging from no noise to only noise.

## 1.3 Dissertation Structure

This Dissertation Is organised Into the following Chapters:

**Chapter 1** is the introduction, which sets the stage for the research and outlines its importance

**Chapter 2** is the background and Literature Review. This section Explains briefly about the Whisper model and its architecture, as well as details in the Whisper paper by OpenAI, before explaining bias in ASR, the socio-linguistic causes of bias in ASR and finally discusses previous studies relevant to ASR bias.

**Chapter 3** This chapter covers the datasets used to test the Whisper model, explaining the criteria for selecting the datasets, noting the datasets used and their features, and briefly explaining any preprocessing completed on the datasets

**Chapter 4** This is the methoodology section, and briefly explains concepts, ideas and tools needed in order to understand the work completed in the results and analysis section. This includes metrics for qualification, processing required, language datasets used, as well as a further explanation of features of Whisper as they relate to the testing in the research.

**Chapter 5** lays out the experiments and their results, as well as short analysis of the results.

**Chapter 6** briefly summarises the main conclusions and findings of this research.

# Chapter 2

# Background

This section intends to cover material written that is either critical to the understanding of the material covered in this dissertation, as well as going over previous literature covered in the development of this dissertation that influenced the direction of the research.

## 2.1 Large Automatic Speech Recognition Systems

ASR is the technology through which machines convert spoken language into text. Through usage in voice assistants and smartphones their adoption increases accessibility and productivity for users. These systems are typically designed with a broad approach, utilising machine learning techniques that involve using large amounts of training data to increase their accuracy and performance. The larger and more varied the training data, the more variability the model will be able to deal with during operation.

ASR Systems were invented in the 50s [12] but more progress was made during the 70s and 80s with the adoption of Hidden Markov Models[10] and the use of dynamic time warping,[27] which allowed the systems to work on much more complicated examples of speech. In the 21st century, the integration of neural networks lead to a large leap in ASR capabilities, increasing the recorded accuracy of the systems in speech-to-text conversion. More recent ASR systems, including Whisper, have reached a level of technical sophistication that allows them to process even noisy input data accurately, and often can successfully deal with a range of linguistic inputs, and to simultaneously transcribe and translate.

### 2.1.1 Whisper

Whisper[23] is OpenAI's ASR system, designed to transcribe spoken word into text with high precision. Through a variety of techniques it aims to be robust in its ability to translate different accents and dialects while also being able to work in many languages. Whisper's versatility is reflected in its performance when compared against Commercial ASR models: It is able to get more accurate transcriptions when transcribing noisy audio, dealing with multi-language transcriptions, and in long form transcription.

This is the result of a very large multilingual dataset pulled from many sources. The significant size (over 600,000 hours of audio and its transcriptions) of Whisper's dataset was achieved by moving away from gold standard datasets and leveraging large scale weakly supervised data; that is data that is filtered for quality but not manually checked. For example, a large amount of data was gathered by pulling audio from online video hosting websites along with its subtitles. An advantage of this source for Whisper is that it provides a diverse dataset covering a range of environments, microphone qualities, accents. This data can be of dubious quality, so these audio files are filtered with a variety of heuristics to remove transcripts that may be of poor quality, or be the output of other ASR systems. The enormous size of Whisper's data means that it has been trained on data that contains more nuances and patterns of speech that may be missed in smaller datasets used to train other models.

OpenAI's approach to training Whisper is intended to increase the robustness and generalization ability of Whisper by exposing it to a range of spoken language scenarios. The effect of this is that Whisper achieves high proficiency when dealing with unfamiliar contexts and situations, such that would come up during real world utilisation.

#### 2.1.1.1 Architecture of Whisper

Whisper is built on neural network architecture, specifically the sequence-to-sequence learning architecture that makes use of transformers. The model consists of an encoder to process the input and a decoder to generate the output, which both use a self-attention mechanism. In Whisper, the encoder and decoder transformer blocks are of the same width, and the number of encoder blocks is the same as the number of decoder blocks.

A Transformer block in the Whisper model is a unit that has 2 components: the self-attention mechanism and the Multi-layer Perceptron (MLP), a position-wise feed-forward network. The self-attention mechanisms allow each position in the input sequence to attend to all positions in the previous layer of the model, and generates a set of attention scores. These scores determine how much focus each element of the sequence has on other elements, allowing the model to determine the important features for completing the task. This is then passed to the MLP that applies a transformation to it, followed by layer normalisation, before this is then passed to the next transformer block. This structure is useful because it allows the model to process complex sequential inputs efficiently. It is important to note that the above Architecture is commonly used in ASR and OpenAI has specified that it has used an off-the-shelf implementation of an ASR in the development of Whisper in order to highlight that the improvements seen are due to the changes in the training as opposed to a breakthrough in model improvements.

For the encoder, Whisper uses audio data that is resampled in 16000Hz and a 80 channel log-magnitude Mel spectrogram representations is used. This allows the audio to be easily broken into frames and extracted for information that approaches human auditory perception. It is then passed into a set of convolutional layers that process the input. A series of transformer blocks then apply the self-attention mechanism to capture insights from the data.

The decoder is tasked with generating the textual output from the audio processed in the

input. It uses cross-attention mechanisms that allow it to attend to the relevant parts of the audio while generating the transcriptions. This is set up in such a way that it can do all the tasks that Whisper is equipped to deal with, such as transcription, translation, and dealing with speechless recording, using tokens to represent languages and tasks. This means that the model is able to achieve what previously would have required multiple task-specific models, improving the generalisation and utility of the Whisper model.

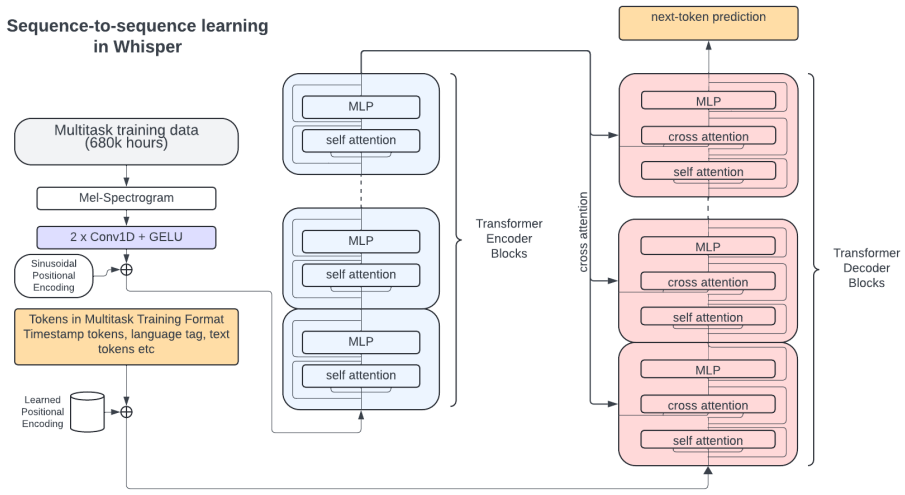



Figure 2.1: Overview of Whisper Architecture

#### 2.1.1.2 Performance tests in the Whisper paper

The model demonstrates strong performance on various speech recognition tasks, including multilingual recognition, translation and more, without the need for fine-tuning for specific tasks. Their evaluation revealed that the Whisper model compares favourably in statistics such as Word Error Rate (WER) against other models, and in fact begins to approach human accuracy on certain tasks. For multilingual speech recognition, Whisper also performed well, but came short in a few translation tasks for more obscure languages, which the paper ascribes to poorly labeled and limited or mislabelled data in the dataset for those languages. The model also demonstrated ability to transcribe long-form audio successfully, utilising a strategy of buffered transcription that eliminated the increase of errors. In this task too, Whisper's performance was seen to approach the results of professional human transcription.

Throughout their evaluation, the authors use a large and broad selection of existing datasets to evaluate the model and demonstrate its ability to cross generalise across domains including multiple datasets that highlight a greater WER for specific groups over others, in what could be interpreted to be bias from Whisper. The datasets used had variety, ranging from very standard delivered English in the TED-LIUM3[7] corpus which consists of the audio from TED talks, datasets like Earnings 21[25] and 22[26] which are datasets of English call data that purports to be provide a greater representation of accented speech, as well as the dataset it performs the most poorly in, the Corpus of Regional African American Language (CORAAL)[11] which is a large corpus of

Regional African American speech. Represented in their results is the understanding that while Whisper does in fact perform much better than its contemporary models, there is still a discrepancy between the transcription qualities for certain varieties of English over others.

As it is relevant to the work completed in this dissertation I also highlight an additional experiment conducted on the Whisper model in the Whisper paper. The Whisper model was tested on noisy audio files to test the ability of the system, and found that it often faltered at a lower Signal-To-Noise(SNR) ratio than other comparable models. This is to say that Whisper is able to transcribe from audio with more noise than other models before seeing the quality of the transcriptions degrade. OpenAI tested Whisper using English from the LibriSpeech[9] corpus, and found that this was true on both white noise and pub noise (noise designed to simulate public space). This allows an analysis of Whisper performance at a wider variety of noise levels than for other ASR systems. Something that warrants further exploration is on the Bias in Whisper models at high noise, as this test does not show the variance that could be present in the at high noise for different dialects and Lingustic backgrounds.

## 2.2 Bias in ASR

Bias in ASR is a complex issue that must be explored to have a complete understanding of ASR and the potential impacts of its adoption . In order to understand bias in ASR, we need to understand how bias can manifest itself and what harms will occur when that occurs. Within ASR systems used commercially, it has already been noted[32] that Youtube's ASR system showed higher error rates for women and African American speakers as compared to their male, Caucasian counterparts. This type of bias can be observed when the data a system is trained on does not represent the whole of its potential user base, leading to unequal performance of different user groups. There is also linguistic bias where the data may represent the average speaker and their language, dialect and accent, but doesn't include much about speakers who deviate from this. An ASR with this kind of Bias will have trouble or fail to work with speakers of these nonstandard variants of English as it has not been trained to deal with them, reducing the ability of individuals to use the ASR system.

The real world implications for the failure to have suitable development sets is the manifestation of a performance disparity that depending on the implementation of the ASR system could have disastrous consequences. For example, any healthcare system that would implement ASR to automate a process would perpetuate the same systemic inequalities on its patients as were represented in the recording of the dataset, treating speakers of non standard varieties of English with less efficiency. Another example could be a hiring system that utilises ASR, which could fail to correctly transcribe due to the candidates linguistic background, causing a difference of outcome for individuals for reasons outside of their control. Fixing this would require the diversification of training datasets or implementing some form of algorithmic fairness measure to stem the bias.

There can also be transcription bias, which while also caused by an imperfect dataset,

can cause issues not in the understanding of the audio, but in the creation of the transcription. For example, if the transcription does not include data of certain forms of speech or colloquialisms, the transcription may find that sequence of words unlikely and consistently decide on an incorrect transcription. This can also only be solved by improving the means with which data is gathered, which would improve the variety and accuracy of transcriptions.

## 2.3 Related Work in Bias in ASR

It is important to consider that this dissertation adds to a wealth of literature that already exists that explores the space of Bias in ASR systems, many of which demonstrates the bias that common ASR systems display when used on speakers of non-standard English.

### 2.3.1 Explorations on the cause of ASR Bias

Bias apparent in ASR systems does not exist in a vacuum. While many of the causes of bias in an ASR can be explained as being due to issues of the diversity of a training corpus, it is important must consider how these biased datasets came to be. In Nina Markl's paper, Language variation and algorithmic bias[15], Markl argues that ASR technologies and other speech and language technologies (SLT) reproduce structural oppressions in the world. The paper means to use a socio-linguistic lens to analyse why there is a difference in the performance of ASR systems in the transcription of groups such as second language English speakers.

The Paper explains that SLTs, especially those based on machine learning, often have a predictive bias that systematically generates higher error rates for marginalised groups, which has a slew of bad resultant outcomes. Markl links this with the unequal access that smaller language groups or communities have to SLT's. This lack of access further results in a lack of SLT resources for said communities and languages. The resultant scenario is where group representation in SLT resources is incredibly skewed away from smaller and more marginalised groups while larger dominant communities such as European languages have a wealth of SLT resources. These differences can be seen to often point towards broader social issues, such as colonial legacies and global power structures.

The study also performs Tests on 2 commercial ASR systems, Google's Speech-to-Text, and Amazon Transcribe. It uses the George Mason University (GMU) Speech Accent Archive (SAA)[35], a database of speakers from a range of linguistic backgrounds, as well as the Intonational Variation in English (IViE) corpus [6]. The focus of this section of the paper is to demonstrate a clear representation of bias within ASR systems by testing for Accents of 1st and 2nd language English speakers. The results demonstrate this bias, with the best performance of the ASR systems seen for "prestigious" varieties of English such as Received Pronunciation, while much higher WER values were seen for second language speakers as well as speakers of regional dialects of English, particularly those from the north of England and Northern Ireland. The results seem to echo broader societal and linguistic hierarchies, and suggest that certain varieties of English,

particularly those associated with marginalised communities may be underrepresented within the training datasets.

Markl concludes that In order to mitigate these biases, it would make sense to shift the focus of work in SLT's towards the experience of marginalised users; As Markl notes, ASR systems are affected by the socio-linguistic biases in SLR resources, and cyclically the affected ASR systems also propagate those same socio-linguistic biases back into the real world. Thus, reducing the harm caused by ASR systems is critical in order to prevent the propagation of these attitudes in ASR applications.

Markl also Cites in her paper Blodgett et al's[2] Paper, a survey of historic papers on Bias in NLP. In it, The paper concludes that in spite of a wealth of Scientific literature existing in the field of NLP, a glaring flaw is present in these piece in the form of a lack of a willingness to engage with the socio-linguistic contexts and causes of these biases and choosing to divorce the topic of the paper from these root causes. We agree with this perspective, and feel a holistic understanding of Bias is required to Analyse bias in ASR systems.

### 2.3.2 Selection of descriptive work on ASR bias

In Feng et Al 's[5] paper, Researchers attempt to quantify and examine the bias within a Dutch ASR system. The paper explores this using the Dutch spoken Corpus[18] and its extension, the Jasmin Corpus[3], as their training and evaluation datasets. These include hundreds of hours in recordings from speakers from the Netherlands and Flanders that aims to represent speakers of a variety of ages, speaking styles, and fluency within the Dutch language. This allows them comprehensively analysis of the ability of the ASR across large speaker demographics.

The paper qualified the bias in the ASR system by measuring the WER of the model when working with audio from the different demographics, as well as a more granular phoneme-level error analysis which would enable the researchers to not only note where the machine would fail, but also be able to figure out which features of a speaking style or accent represented the largest obstacle for the ASR system.

The study found that the ASR system did seem to exhibit some bias against certain groups, as there was a significant difference in the WER of the transcription across the demographics. For example, there was a much greater WER when dealing with speakers younger than the age of 30, as compared to their older cohort, as well as having difficulty transcribing the words of less native speakers of Dutch. The paper notes that there are some ways with which Bias should be aimed to be mitigated, starting with the diversification of datasets, but notes that it should be an active process for the developers of an ASR system to regularly analyse the performance of the ASR across the breadth of potential user demographics, and then on the insights taken from this analysis address the biases by implementing corrective measures (such as retraining).

Another paper by Koenecke et al[13] assesses the performance of the ASR software of 5 major leading companies and compares their abilities to transcribe speech from African American and Caucasian speakers. This research used 2 key datasets, the CORAAL[11] corpus and Voices of California(VOC)[30] comprising of interviews of diverse regions

and backgrounds, and made sure that they had a metric for the presence of African American Vernacular English (AAVE), within each audio snippet, in the form of a dialect density measure, in order to better pinpoint the source of the problems in the ASR's performance

The findings of this paper revealed that the ASRs underperformed for black speakers by almost twice as much as it did for white speakers, with a WER of 0.35 and 0.19 on average respectively. This discrepancy in the ability of the models to transcribe speech was especially pronounced in test where the phrase spoken by the black American and the white American were identical. The researchers aimed to explore the potential source of the disparity by examining the components of the ASRs individually, and found that the primary source for the disparity was in the acoustic model, which struggled with the phonological and phoenetic details of AAVE, rather than the grammatical or lexical characteristics, which the researchers concluded could have been brought about by a lack of data on AAVE being present in the training data for these ASRs.

Martin and Tang[16] Demonstrate in their paper a more microscopic and focused look at a specific linguistic feature within dialectic English in their paper. Previous papers have established that there is an apparent bias against African American speech in ASR systems, and in this paper they seek to looks further, singling out the Habitual 'Be' - an invariant form of be that represents a habitual aspect - and analysing how well ASR systems are able to deal with them. Their paper uses 2 commercial ASR systems, and used the CORAAL dataset for their purposes, tagging instances of 'be' in the corpus to signify if they were examples of the habitual 'be'. The Study was able to find evidence that there was an increased error rate in transcriptions for not only the the instance of the habitual 'be', but for words in the immediate proximity of it.

In their analysis, Martin and Tang note that they were able to control for most other factors, suggesting that the core issue lies with the ASR model. This demonstrates that the bias in the tested ASR systems is not limited to difficulty with phoenetic and prosodic elements of dialect but grammatical dialectical variations also appear to be difficult for ASR systems.

In the practical experimentation further on in the paper, the Edinburgh International Accents of English Corpus (EdAcc)[1] is used as one of the datasets for testing the abilities of the Whisper model. In its attached paper by Sanabria et al, as a analysis of whether the corpus succeeds in being a more more diverse dataset, as well as proving its utility as a tool for testing bias in ASR by proving that a variety of ASR systems do in fact have a noticable difference in WER between different first language speakers. Tests were conducted using Whisper, a anonymised Commercial model and a third model, Wav2Vec2.0. Of the 3 models, Wav2Vec2.0 performed the poorest, with Whisper being the most advanced of the three and performing the best, thus demonstrating the improvement seen from using large weakly supervised training.

Given that I am using the provided corpus and performing an analysis of Whisper, its important to consider what has already been tested by the EdAcc team in the paper so that I can expand on it. As the topic of the paper focuses more on the creation of a new corpus, they do not complete too much testing on the dataset, as the testing is completed to represents a proof of the ability of their dataset to reveal bias in an ASR, as

opposed to a full investigation of Bias in ASR generally. Given the dyadic nature of the conversations in the EdAcc, the paper opts to only perform WER calculations on samples where both speakers have a similar linguistic background. This not only reduced the number of tested files, reducing the statistical value of the results, but also decreased the reliability of the results, as there may be other non-controlled factors between the speakers that would result in larger or smaller WERs. The paper acknowledges this, as well as other shortcomings, in their analysis of the results. From their results, it is easy to say that the EdAcc corpus may not be sufficient on its own for the analysis of bias in Whisper. The number of speakers in the corpus in each of the categories it identifies is too small and may be more easily swayed by tertiary factors for individual speakers within the categories. For this reason, the results read as unreliable. While this corpus will be used in this dataset, past work using the dataset leads us to believe that it would be most useful when not used as a sole metric for the analysis of bias.

While these other papers often are working on older and less contemporary ASR systems than Whisper, and consequently may have been trained on a less varied and less representative training dataset, its important to consider the work done so that we can contextually understand the variety of ways that bias can manifests itself in ASR systems, the socio-linguistic factors to keep in mind when discussing bias. They also acts as a benchmark point from where we can examine the improvement that Whisper has made relative to these older models in using a diverse and less biased training set.

The work completed in prior work is also useful in the ways that it illustrates metrics that can be used to analyse Whisper alongside methods to present the results with high clarity. On top of this, past work done in the same field gives us something to base our expectations when it comes to Bias as it presents itself in ASR systems.The work above broadly seems to indicate that ASR systems have historically greatly skewed results towards speakers of 'standard' and privileged varieties of English, while having decreased performance when tested against speakers of out-groups of many varieties, such as minorities and regional populations that are socioeconomically disadvantaged compared to other groups. While Whisper's own work seems to indicate that their model performs better in these metrics than previous ASR systems, a degree of bias can still be seen in their results. In this paper we will expand on OpenAI's results to explore the breadth of bias in Whisper.

One problem I feel with some of the past work is that due to some results having groupings of speakers with filesizes somewhere in the low double digits. While the results they draw are still useful, as I read the EdAcc paper [1], or the Martin/Tang [**?** ] I cannot help but find the results slightly unconvincing due to the low speaker counts and the use of just a single corpus. Experiencing this, we sought to use at least 2 corpora, ensuring that one was of substantial size

Additionally the above work highlights ways that work in the exploration of bias in Very large ASR systems can be expanded in the future, with more microscopic analysis of features within languages being potential ways that further investigation beyond the scope of this dissertation could look at,

# Chapter 3

# Data

This section introduces the datasets used in this project, as well as the methodology of their gathering and processing prior to transcription.

## 3.1 Datasets Selection

In this dissertation 2 datasets are used for the testing of Whisper. There were a variety of Criteria for the selection of a corpus, which are the following:

- The dataset must have a high quality of audio. Low quality audio can affect WER, and noise will be introduced in a controlled fashion to test the transcription of dialects so the audio recordings must be of a sufficient quality.
- The dataset must by design represent a diverse demographic and dialect spread of speakers, down to features such as linguistic background and place of origin.
- the data must have high quality Metadata.
- the dataset should be from a broad scenario that contains many different facets of speech, so that many elements of the dialect or language can be analysed or be statistically represented.
- the dataset should represent a very large number of speakers, ideally many from each dialect or linguistic background in order that that data gained from the background is statistically significant. This is in order to preserve the reliability of the results
- Understanding the limitations of my abilities as a Undergraduate student, the Corpus must not be so big as to be unwieldy and difficult to test with.

The above conditions in a single corpus would be incredibly difficult to curate, and thus it is very difficult to find a corpus that satisfies them all. We aimed to find more than one corpus that could be used to test for different scenarios, and therefore get a high coverage of criteria. Ultimately, We decided to have one broad short form English dataset that would allow us to study a large range of speakers without massively increasing the operational load in order to get transcriptions (SAA). We could also

also supplement this with a separate longer-form English corpus that could due to the specificity of the recordings, reveal failings of the Whisper model that the shorter corpus would miss (EdAcc).

A variety of other corpus were considered, such as LibriSpeech[21] or the TED-LIUM[7] Corpus, amongst others. However, Many of these corpus were very large and unwieldy, going above 1000 hours of audio. Other corpuses were like the EdAcc, in that they were corpuses that contained 50-100 speakers, but often did not represent great variety in the linguistic features of the speakers. Another factor that contributed to the corpora chosen were issues of secondary details about the speaker. Many datasets did not include tables of information about each speaker that would be useful to better diagnose bias detected in the testing.

## 3.2 Edinburgh International Accents of English Corpus

The Edinburgh International Accents of English Corpus[1] is a dataset consisting of over 40 hours of speech, with 121 speakers of a variety of linguistic backgrounds speaking conversationally, designed to be a better representation of speech that datasets before it. The dataset also includes a table of detailed information about the speakers, such as linguistic information, but also includes factors about their background, such as accent self-identification. The strengths of this dataset are numerous. For one, by using speech recordings of casual conversations, the dataset includes dialect features of speech that are only seen in the real conversation and missed in more rigidly formatted datasets. The Dataset has been specifically curated to solve the problem of lack of variety in current ASR datasets meaning that it is a good fit for my tests; as mentioned in the background, in their own testing they identified that due to the greater diversity of their speech their corpus was better able to highlight bias in ASR systems when compared to contemporary datasets.

The Transcription of the EdAcc dataset was completed by professional transcribers. This means that the transcription data is of a high standard: every turn in the conversation was manually timestamped and segmented, with multiple non-speech features labelled in the transcriptions, such as speaker overlaps, laughter and others. The high quality of the transcription allows fast and accurate analysis when comparing against Whisper generated transcripts.

### 3.2.1 Preprocessing

As this audio was entirely recorded from dyadic conversations, there were 2 speakers to each recording. In order to use the audio to test the ability of Whisper against the speech of a given individual, it is necessary to split each piece of audio into 2 separate audio files, isolating the speakers. The database provided a timestamp of every turn in the conversations, and based on the timestamps the audio was split into its segments and then composed into a file for each speaker. Kaldi[22] notation is used in the transcriptions to denote whenever the speech is unintelligible due to noise, or when individuals speak over each other. Therefore, we elected to remove all segments with this notation in its transcript. The resultant audio segments are a continuous intelligible

recordings of a single speaker who's features are recorded in the speaker table, and are ready for transcription. This is not perfect, as the Gold standard transcriptions do not have perfect timestamps for turns. One way this manifests itself is some of the processed audio still containing speakers speaking over each other.

## 3.3 GMU Speech Accent Archive

The speech accent archive (SAA)[35] is an online resource of over 3000 recordings of speakers with a variety of linguistic backgrounds created to document the diversity of accents in the English language. It lists each recording of a speaker alongside information such as the speakers first language, and history of speaking English. These files are hosted on a website arranged in a directory tree, where the directories are the primary language of the speaker. Every recording is a reading of the same elicitation paragraph designed to contain a variety of phonetic elements and sound combinations to highlight differences in pronunciations between accents. The ease of comparing L1 and L2 speakers with this dataset inspired the direction of the research.

A strength of this dataset is that it is easy to compare side by side results from one speaker or group to another due to the standardised elicitation. Additionally this resource has a very wide breadth, encompassing over 222 language backgrounds, providing ample recording counts for each language. However, due to its standardised format, it has some weaknesses, such as dialect specific words and phrases not being able to be tested. Additionally, Unlike the EDACC corpus, the recordings do not have manually verified transcriptions, and sometimes the speakers will make small mistakes when reading the paragraph, leading to a slightly inflated WER. As there are many datapoints, these have a minimal impact on WER values across languages.

The total length of this dataset is approximately 24 hours of speech, spoken by 3032 speakers, which is more individual speakers than Librispeech[21]. This Dataset has been previously used in other papers [15], However, the corpus has increased by more than 1000 speakers since its previous use and now represents a broader corpus than past results.

### 3.3.1 Retrieval and Preprocessing

As this resource appears to be mainly for linguistics students, there is no readily available way to download all of the dataset as well as all of the metadata attached to each speaker. The search function is also currently unavailable, and therefore gathering all of this data required the creation of a rudimentary scraper that utilised the BeautifulSoup[24] python library to go to every file in every language directory on the website and download each audio file individually. We also scraped the information about the speakers from text blocks in the margins.

# Chapter 4

# Methodology

## 4.1 Whisper

As previously introduced in the background, the Whisper model is an ASR system that uses an very large training set to achieve a greater robustness than other ASR systems. In this chapter, we cover elements of the Whisper model that are directly involved in the testing.

### 4.1.1 Models

Up to this point I have referred to Whisper as a model, but Whisper is also the name of a family of models of all different sizes, designed to show the scaling properties of Whisper. As larger models are used for transcription, the amount of time required by the model, The size of the model on disk and the required VRAM increases. The specifics of the structure of the models are as below Note that in addition to the models in the

| Model | Layers | Width | Heads | Parameters |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39M |
| Base | 6 | 512 | 8 | 74M |
| Small | 12 | 768 | 12 | 244M |
| Medium | 24 | 1024 | 16 | 769M |
| Large | 32 | 1280 | 20 | 1550M |

Table 4.1: Model Sizes and their Architectural Details

above table, Whisper also releases English language only versions of all of their models below their large model, which only accept English language Audio. They perform slightly better at these tasks at the tiny and base level, but perform similarly for the larger models. These models were considered for testing, but it was concluded that this would most likely give relatively uninteresting results relative the amount of time taken to run and process the data from these models. As well as the Large model in the above table, Whisper has since the public release of Whisper produced large-v2[19] and

large-v3[20], newer models that promise better transcriptions. However, with practical considerations I decided to test results on data transcribed mainly with the 'base' option, primarily as at the scale of an undergraduate project, this allows for transcription within a reasonable time frame. I also completed some tests on all models up to the medium size in order to analyse if there was a great variation of bias between the different model sizes, and if the smaller models suffered from more bias than would be expected. I did not complete an analysis of any of the models of size large and above. This is because of the VRAM load and the time for transcription required. The medium model run on the GMU SAA required 1.5 Gigabytes of space on disk as well as 14 hours of transcription time, coupled with a noticeable performance drop of the device.

The drawback of this is that while many previous papers do not specify the model size of Whisper used for their work, they will have likely used larger models, preventing direct comparison between results in this dissertation and others. Additionally, it will not demonstrate the best performance that Whisper is able to give. We acknowledges these shortcomings, but still contends that the results analysed in this dissertation are still useful. Given our findings using the Medium model, we believe that scaling up the model will still reflect the same biases, albeit to a smaller degree.

### 4.1.2 Language in the Whisper Model

Whisper, as covered in the background, is able to do more than transcribe English audio. Whisper also includes the ability to transcribe audio from a variety of other languages into English. While the scope of this dissertation is exclusively the bias demonstrated by Whisper within the confines of the English language, due to the above feature the Whisper API allows a user to pass Whisper the language of the audio it is transcribing. In most of our testing, this was always set to English. However, in lieu of being passed a language Whisper will instead take the first 30 seconds of audio from a clip, pass the Mel-spectrogram of this snippet of audio to a language detection unit that detects for the presence of language tokens. It will then use these tokens to calculate the probability of a given language being the language of the audio. It does this for all 99 languages Whisper is compatible with, returning a dictionary of probabilities and selecting the most likely of these options. Whisper will then transcribe if the language detected is English, and will otherwise attempt to Translate from the detected language. In this paper this is treated as a possible vector for Bias within the Whisper model, and will be analysed alongside the transcription functionality of Whisper.

Although outputting this detected language is not a part of Whisper API, one can easily receive an dictionary of predicted languages as well as each languages assigned probability score by using lower level methods within Whisper.

It should be noted that Whisper represents its languages using the International Organisation for Standardisation (ISO) 639[9] set 1 codes, which are 2 letter standardised codes, which need to be converted to set 3 codes in order to work with the language sets used in this experiment

## 4.2 Languages

The SAA alone has speakers listing 227 distinct languages as their primary language. The number of distinct languages listed as a first or second language for speakers in the SAA is undoubtably higher. While a larger language breadth does allow a more specific analysis of what languages do and do not perform well within Whisper, it is difficult to ascertain broader trends with so many discrete categories.

This introduces the need to find a way to classify or sort languages into groups. This can be done with a variety of metrics such as common linguistic origin or ancestor and feature similarities, and these features can often be found in Language Structure Databases. However, while these databases can often provide similarity scores between different languages or create language clusters for classification, there are often disagreements between linguists on the most important factors or the correct way to classify this similarity, and so any groupings or trends gleamed using these databases may be specific to that database.

One strategy that was attempted in our work was the use of the database at the Automated Similarity Judgement Program (ASJP)[36]. This is a database that aims to compile a list of 40 specific words from all of the worlds languages. The information from these wordlists is then used to identify linguistic features of said language. This is a enormous database that contains almost 6000 languages. In many papers[28] that utilise the ASJP, a metric known as Levenshtein Distance is used to calculate the distance between 2 languages. However,the opinions[8] of some experts within the linguistic fields is that this database is neither useful nor adequate for the classification of languages due to it ignoring elements such as suprasegmental traits. Additionally, As my project moved away from using lexical distance as a metric, the value of the ASJP was greatly diminished, and ultimately it was not used.

The second strategy looked at was the use of the Word Atlas of Language Structures (WALS)[4] database. this is another database that features thousands of languages from around the world aimed at helping typological analysis and expressing linguistic diversity. Compared to the ASJP database, the WALS database is stronger as it contains much more broad coverage of features, including language details such as their syntax, morphology, phonology, lexicon and family making it a very detailed resource for comparing languages. Again, due to conflicting opinions about the correct or most useful way of classifying languages, there is some debate about the utility and accuracy of this database. However, I believe that for my needs purely as a classifier of languages origins this database suits the needs of this paper.

#### 4.2.0.1 Using WALS

Initially The method of classification I decided to use within my project work would be to create KMeans clusters using all of the language features in the dataset. This would allow me to identify clusters of languages with similar features that had increased WER values, which would represent Whispers poorer performance on languages with thees features However, it was decided that this would falls slightly outside of the scope of an undergraduate level dissertation, given the amount of data processed in this

project. This could be a potential area for further study in a future project. Instead, this project has opted to mainly use linguistic families and genera to aggregate information from a spread of linguistic backgrounds, with super-regional information or coordinate information also used.

It is worth noting the shortcomings of the WALS database. Unfortunately, as there are gaps between the specificity of linguists and laypeople, there are a few languages in the corpora that are unclassifiable, although I believe that the rest of the project is wide enough in scope for our findings to still be of value. An example of this is the Mongolian language. While there are many 'mongolian' speakers in the SAA , the Mongolian language with the ISO code "mon" is not in the database, as WALS and Ethnologue[33] considers it a Macrolanguage, and therefore the self identified 'Mongolian' speakers could be speaking a range of languages. These records are therefore unclassifiable. Additionally, smaller languages, such as those found on the west coast of Africa, or varieties of languages spoken in New Guinea were are victims of obscurity begetting a lack of information and could not be found in the WALS database. The result of this is that about 132 files are unclassifiable without a higher order of work needed to manually classify them. This gives us a classification rate of 95.0% for the files in the SAA. Of the unclassified languages, not one language had more than 10 recordings of unique speakers. Due to these factors, we decided that it would not be worth chasing the specificity of the results that could be attained with these rarer languages, as they would require a large time investment for very little added value to our results.

With our testing, we found that the WER of members of larger groups such as language families or genus tended to be quite similar to each other. This meant that often times, we could gain the results from a super-groups and use that information to make assumptions on the performance of Whisper on a more specific language that tended to be correct. an example would be of Semitic languages which had a WER of 9.8%, and Arabic, which had a WER of 9.3%. Consequently, in order to reduce the number of categories worked with, we will show results of a genus, or family, rather than an individual language, based on discretion on whether the values for the group are similar to the values of any member of the group.

Finally, In spite of our best efforts, While the linguistic specificity to refer to specific dialects of foreign languages exists within the WALS database exists, such as the division of Arabic into a wealth of subcategories, we have decided to group together first language speakers of dialects of other languages together. Although the mutual intelligibility and divergence of these languages definitely affects factors such as WER, our analysis already contains roughly more than 220 languages, and so increasing the linguistic categories risks making the data harder to interpret. Therefore, outside of English, Languages are treated as single unified groups.

### 4.2.1 Noise Generation and Mixing

In order to test the quality of the transcriptions under noise, we need to first create some noise to mix with the original audio. The generation of white noise is relatively simple. For one second of noise, using the default sample rate of 44.1kHz, 44100 values will need to be created. A value is generated from a normal distribution centered around

zero with a standard deviation of one for every sample. In order for this number to represent an amplitude, it is normalized so that all values sit within the range of -1 and 1. This is then converted to 16bit PCM format that can then be written to a desired audio format, whether that is .MP3 or .WAV. We then mix this into the main audio.

The measure that we use when mixing the noise and the base audio is Signal-to-Noise Ratio (SNR). SNR, expressed in decibels, is a useful measure for audio processing that compares the power of the signal against the power of the noise. The larger the SNR, the quieter the noise. It is normally used when testing audio recordings, but here it allows us to increase the level of noise in a controlled manner.

Mixing is achieved by firstly calculating the root mean square power of both of the audio files in order to quantify their levels. This is because the audio taken, either from the EdAcc or the SAA are not standardised in terms of equipment used for recording, and therefore will be of varying volume. Using these values and given a SNR setting, the volume of the noise is adjusted to the desired prominence relative to the base signal, resulting in a continuous and controlled white noise level over the top of each file. This is done for every file in the corpus, with SNR ranging from ∞ SNR, representing no noise, 40 SNR, and then in intervals of -10, down to -10 SNR, at which point the original audio is completely drowned out by the noise signal.

## 4.3 Normalisation

Initially, the normalisation of the transcripts was handled with code written specifically for the project. However, due to the innumerous edge cases including special character transcriptions and the transcription of non word elements, our normalisation was not of sufficient quality. We determined that time taken to develop a normalisation function would be better spent in designing the experiments, and consequently for this dissertation, the python library jiwer[33] is used for both normalisation and calculation of WER.

jiwer is a commonly used tool in ASR research that can provide a variety of metrics for the quality of a transcriptions. jiwer has a robust normalisation process that can be flexibly modified with rules and custom standardisations. For the normalisation of the texts in this project, both the ASR output and the golden standard transcription will be converted to lower case and stripped, as is standard, but additional jiwer transformations that were found to improve the WER across the board were also used. For example, jiwer.RemoveKaldiNonWords was found to reduce the WER of all Whisper models on the EdAcc files, and so was added to the applied transformations, but as we found that jiwer.ExpandCommonEnglishContractions seemed to increase the Word error rate across the board, it was not used.

On the EdAcc Corpus, we compared the quality of our normalisation against the normalisation performed in the EdAcc Paper. Within the Paper, the Whisper model used (Whisper large) produces a WER of 19.7% across the Corpus. Our normalisation, which was suitable when used on the SAA, produced (Whisper, small) a WER of 23.0% on the same corpus, more than 3 percentage points worse. Although at least some of the difference could be attributed to the difference in model size or imperfect splicing

of audio creating artifacts, but we believed that at least some of the difference was probably caused by difference in normalisation.

We explored the idea of using the exact normalisation as in the paper. The paper used a Transcription Strategy that consisted of the use of hubscr.pl, which is a program from the SCTK toolkit[17] that that is used to score the output of speech recognition software against reference transcriptions, Developed by the National Institute of Standards and Technology. It is widely used in Speech Recognition research. Here it is used to gain valuable metrics such as WER or even Sentence Error Rate. Additionally, the corpus comes with a Global Mapping File (GMF) that contains all rules for normalisation of the texts as well as the removal of paralinguistic features. However, in order to use hubscr.pl, the hypothesis transcriptions must be in a Conversational Time Marked format. Unfortunately for this project, we were unable to convert Whisper text output into this format in order to take advantage of these higher quality normalisations, and parsing the GMF file to extract the rules would require a large amount of time that was better spent elsewhere

The effect of this is that the WER recorded in the EdAcc in this paper may be larger than the true WER values that Whisper is able to output. The implications of this are that this paper will be a poorer quantitative representation of the Bias within Whisper that could be used to compare it to other ASR systems.. However, I believe that this is not as big of a issue as it seems in analysing bias. This inflation of WER is universal and applies to all files transcribed, and so while the values returned by jiwer will be higher, we will still be able to identify difference between the transcription quality of different speakers and linguistic backgrounds, and identify the ways and extent to which Whisper demonstrates bias.

## 4.4 Calculation

Jiwer was also used for the calculation of the WER. As stated before, WER is a metric that represents the accuracy of a transcription, and it does this by measuring the operational distance of 2 sequences of words.this is calculated as follows

$$WordErrorRate = \frac{S+D+I}{N} \tag{4.1}$$

where:

- $S$ is the number of substitutions,
- $D$ is the number of deletions,
- $I$ is the number of insertions,
- $N$ is the total number of words in the reference transcription.

WER is the most common metric for testing and evaluating the performance of ASR systems, and has been used to measure transcription quality in all ASR related papers cited in the background.

While this was suitable for the EdAcc corpus on account of its high standard of transcription where that even non word utterances were transcribed accurately, We found that this formula returned a much larger WER for specific speakers in the SAA dataset, and that this difference in WER could not be accounted for by Whisper. Due to unfamiliarity or discomfort with recording, or reading the elicitation paragraph, some of the speakers in the SAA dataset would occasionally trip over words and then return to repeat it, or use filler words that were contributing to a much larger WER than would be expected. In order to mitigate the impact of this, when dealing with this dataset the Word Error Rate is calculated as the below instead, with the insertions ignored.

$$WordErrorRate = \frac{S+D}{N} \tag{4.2}$$

While this issue cannot be entirely mitigated, as any mistakes made by the speakers are attributed as a transcription error by Whisper, it does help bring scores that were inflated by poor readings more in line with the rest of the dataset, mitigating the impact of some of the above, and as this is applied evenly across the SAA, it should affect all language backgrounds equally. A side effect of this change in metric is that the WER can no longer exceed 100%.

## 4.5 Statistical Tools

This section discusses methods used to determine the statistical significance of results used in this dissertation. These are important as if 2 distributions are not statistically significantly different any claims made from the result are not very credible. All analysis in this paper was completed using methods from the scipy library[34].

Initially, the statistical analysis was completed using t-tests[31], a statistical test that assesses whether the differences between 2 groups are statistically significant assuming that the data is normally distributed. We soon started to encounter unexpectedly high p-values for similar distributions. We tested our distributions using the Shapiro-Wilk test[29] and found that the normal assumption was being violated; due to the WER results lying on a percentage scale, when results were close to zero, the distribution resembled a truncated normal distribution (affecting the variance). Consequently we decided to shift to statistical significance being determined using p-values from a Mann-Whitney U test[14]. Although this test is less statistically efficient, it can test values without a normal assumption. Whenever statistical significance is mentioned in the results, we are verifying that this p value is below 0.05.

## 4.6 Details of Experimental Process

This section covers additional experimental design elements and issues faced during the testing.

### 4.6.1 SAA Analysis

#### 4.6.1.1 Filtering files in the SAA

In order to mitigate the effect that outlier files would cause, files that did not meet specific thresholds were discounted. For most of the tests done, there is a length threshold of 60 seconds. Similarly, a WER value of above 0.5 is also ignored, as although this could be a sign of failure from Whisper, it is more likely to be a Human error from the individual. After this, A Whisper model of each size generates the transcription of the remaining files. As a final step, if a transcription contained a word count of longer than 80 words, it was also removed from the dataset. The length of the stock phrase is 69 characters, and so anything that greatly deviates from this is most likely down to recitation error more than anything. Additionally, There are files in the SAA that are of no value to this analysis that must be removed. These are any files containing recitations by members of the deaf community. They are not analysed in this paper as we felt that it would be difficult to draw any useful conclusions about Whisper when it is working on audio that we have determined cannot be transcribed. It is important to note that only 40 files were removed from the data in the process of this filtering, which was a scale that allowed us to manually evaluate files that were discounted. Often we were able to determine reasons for the unusually high WER that validated the removals, such as speakers repeating sections of the text, or unusually high natural noise levels of the clips. Of languages with at least 10 recordings included, recordings of that language that were removed never constituted more than 10% of the recordings, reducing their impact. While this filtering does not remove all instances of outlier recordings in this in the dataset, this culling hopefully increases the quality of the remaining set.

### 4.6.2 EdAcc Analysis

The EdAcc corpus includes self reported accent values for individuals, with no restrictions on what the individuals are able to input into the Accent field. This causes a few problems.

As I recognise that any labels I may apply to speakers may be biased, I decided that I will not use my judgements to classify any Accent Values. Note that this did not preclude me from making accent umbrella categories (eg ”Scottish (Fife)” and ”Glaswegian” as ”Scottish) However, There are many values in the EdAcc that make the file unusable for the purpose of accent analysis. Firstly, many speakers applied descriptive accent terms that were not very useful for comparison with other accents. Labels incuded ”Fluent”, ”Native” and ”Generic middle class white person”. As frustrating as this was, for the purposes of these tests, these files were removed from the tests. Additionally, categories such as ’American’ were often self reported by speakers who had never lived in the US, and by subjective opinion, may have had American features of accent based on method of language aquisition, but had accents that would not be considered similar to one of an American. Its possible that a confidence in fluency leads people to use ’native’ categories to describe their own accents. I think this is a flaw in the EdAcc corpus that could possibly be improved by reducing uncertainty with what answers are accepted for this category.

# Chapter 5

# Results and analysis

This chapter covers the analysis of the transcriptions generated by the Whisper models, with the aim of observing and analysing bias found in the Whisper ASR system. Over the course of this dissertation, several experiments were carried out with varying goals to cover the questions introduced in the introduction, resulting in the following tests.

- Analysis of the corpora on the 'base' Whisper model
- Analysis of Whisper performance as model sizes are varied
- Analysis of the language detection abilities of Whisper
- Analysis of Whisper performance as noise levels are increased.

## 5.1 Baseline Analysis

This section covers the analysis of the Whisper models of size 'base'. In this analysis, the model informed that the audio is in English, and then proceeds to transcribe the files, outputting the transcripts in a separate directory to be processed.

### 5.1.1 results on the SAA

This section analyses the Whisper model's performance on the SAA corpus. Here are the results of the 'base' model Whisper on all of the languages with at least 10 speakers in the SAA corpus sorted by their average Word error rate.

This table has been included because it illustrates with the least information, that there is a performance difference between speakers of different linguistic backgrounds in this corpus. The left-most high results indicate that there are some groups for whom the Whisper model is consistently poor with, while a contingent of language speakers in the far right of the chart seem to suffer very little mis-transcription from Whisper. Even if speakers above the 95th error percentile and below the 5th error percentile are ignored, the standard deviation of the WER in this dataset is still 5.12% indicating a wide range of transcription quality.

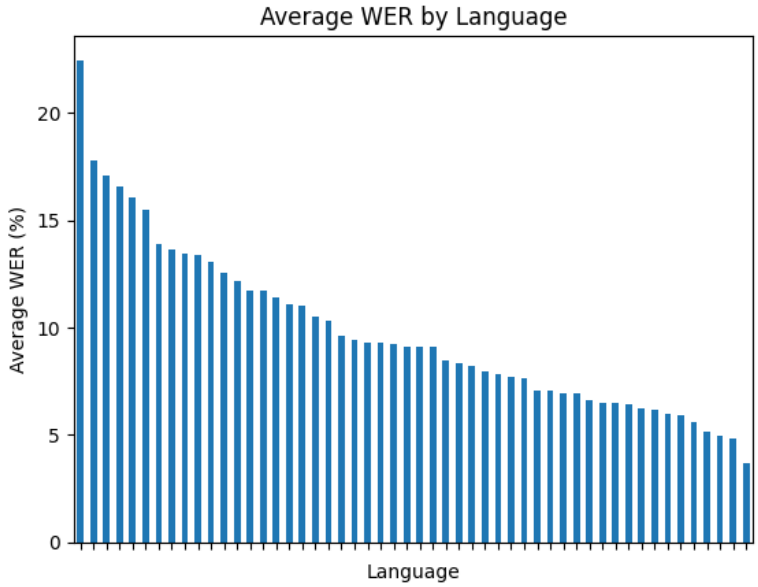



Figure 5.1: Sorted bar graph of the average word error rate for all unique first language in the corpus with more than 10 speakers

Firstly, we broke the data into English speakers and 2nd Language (L2) English speakers, which represented a 660/2204 divide of data samples. We break the results further up in order to analyse the data. For the purposes of this linguistic analysis, as we were dividing the speakers into larger groups based on linguistic classification, we decided to group speakers of divergent pidgins or patois of English into the English group. This is because these varieties, while distinct in their own right, share a significant amount of vocabulary, structure, and historical development with Standard English, making them more closely related to English than to any other language group. In addition to this, it should be assumed in this analysis that any super-group that would otherwise include English, such as Indo-European, or Germanic, will not include English.

Our results found that when we split speakers into separate larger linguistic families, there was a large difference between the transcription quality of L1 speakers as compared to speakers from different linguistic backgrounds.

The table excludes some groups identified in the corpus either due to a comparative lack of recordings from members of said linguistic groups that would have less statistically valid results, or families that act as singleton families, containing only themselves. The latter are analysed further at the genus level exploration. Another choice we made with this table and for others in this section of the analysis are the use of quartile bars to indicate the span of the results. Initially, standard deviation bars were used, but due to the cutoff at zero inherent in percentage distributions, the observed distribution is truncated and consequently non-symmetrical, so the error bars would extend past 0%, making them poor for demonstrating the spread of the WER values.

From the results, It is clear that Whisper struggles with accented speech of L2 speakers much more than that of English speakers, as with the exception of Uralic language speakers, no family averages less than double the WER of the English speakers. While

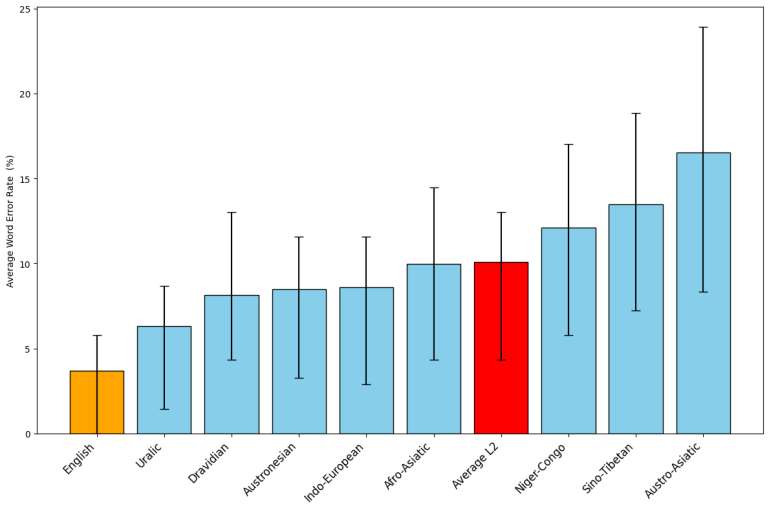



Figure 5.2: Table with Results of average WER by Linguistic Families, With Quartile Bars. The L1 average is in orange, and the L2 average is in red

it is not surprising that languages with great geographic distance from anglophone nations would have higher WERs, it is surprising to see Indo-European languages have a relatively high WER, as one would assume that in the corpus of data for Whisper, L2 speakers with this linguistic background would probably be more greatly represented than speakers of other linguistic background. All groups were statistically significantly different from L1.

These results can be further broken down to the Genus level in order to ascertain the relative performance of more granular language groups on the Whisper model. First, due to the peculiar performance of the Indo-European Language group, as well as their frequency in the corpus (Indo-European represents 1155 individuals, almost half of the non-native English speaking segment), an analysis is done on subgroups within the Family with at least 40 speakers in the dataset.

As the above figure this one also highlights English, non-English and Indo-European values.

This figure reveals some interesting facts about the performance of the Whisper model for speakers of L2 English. Notably, Germanic languages actually have a WER performance that is not hugely dissimilar from the English first language speakers, with a narrower WER difference than any other genus in this test. This is not surprising, as due to the cultural and linguistic proximity of English and German, a German Accent may contain fewer speech features that are uncommon to English Speech and so can be more easily understood by Whisper. Additionally, due to this cultural proximity, the Whisper Model's training data of online videos most likely contains transcription and audio from videos by first language Germanic language speakers speaking in English. This would result in greater performance by Whisper in transcribing speakers with this background.

One element here that is rather surprising is the Performance of Whisper on first language speakers of Romance languages. Of the genera tested, Romance is worse

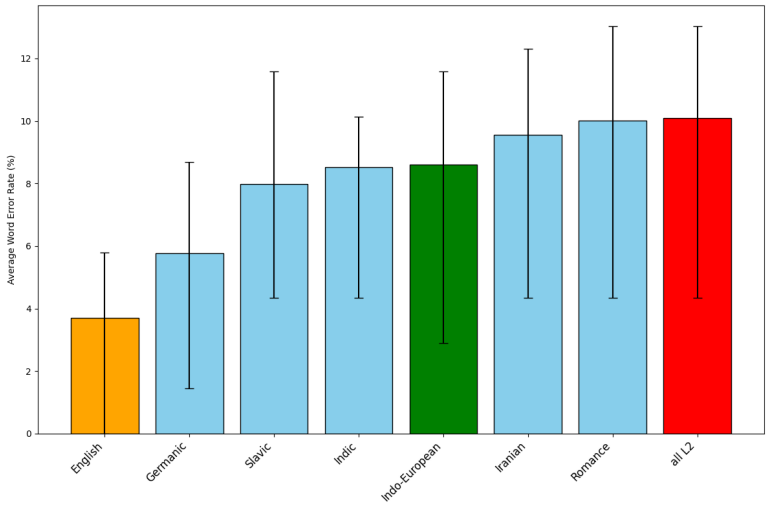



Figure 5.3: Average WER of language genera within the Indo-European Family. L1 speakers are in orange, All Indo-Eurpoean speakers are in green and L2 speakers are in red

than all other Indo-European genera, including languages that would be predicted to perform badly due to lower cultural proximity, such as Iranian, or Indic. Romance is in fact the worst performing Genus in the Family. All distributions here are statistically significantly different from the L1 distribution.

A further analysis was done on the languages within the Romance family in order to explain it, or otherwise to see if it was caused by an outlier language or other. The results are as Below

Table 5.1: Mean WER of Whisper Model for L2 English Speakers from Romance Languages

| | Romance | French | Italian | Portuguese | Spanish | Romanian |
|---|---|---|---|---|---|---|
| Mean WER | 0.10004 | 0.09088 | 0.10507 | 0.08333 | 0.11089 | 0.06401 |

Table 5.2: Mean WER for L2 speakers within the Romance group

Here it can be seen that while the WER for the Romance languages are for the most part not too far from the WER value for Romance Overall,the largest contributor to the high WER can be seen to be from Spanish Speakers. A further look into the data for the speakers reveals that of the 245 distinct Spanish first language speakers who were in this dataset, only 24 were from continental Spain (which with a WER of 12.7% are actually harder for Whisper), with the rest of the Spanish speakers being from Latin America, making the results more representative of that subset of Spanish speakers. This result could be because of Whisper's dataset of online videos lacking videos of spanish L2 English. The internet has a thriving and lucrative Spanish language media sector, so Spanish individuals may not be inclined to reach a English audience. Given that having good transcriptions is mainly done by channels with reason to transcribe (it must be lucrative or affect a large enough sub sector of viewers), then it would therefore

be likely that Whispers training set, while it may contain many hours of Spanish content with English subtitles, contains very little spanish L2 English speakers.

We also did an analysis of some of the genera in the other families, as well as introducing the Korean and Japanese genera. The results are below

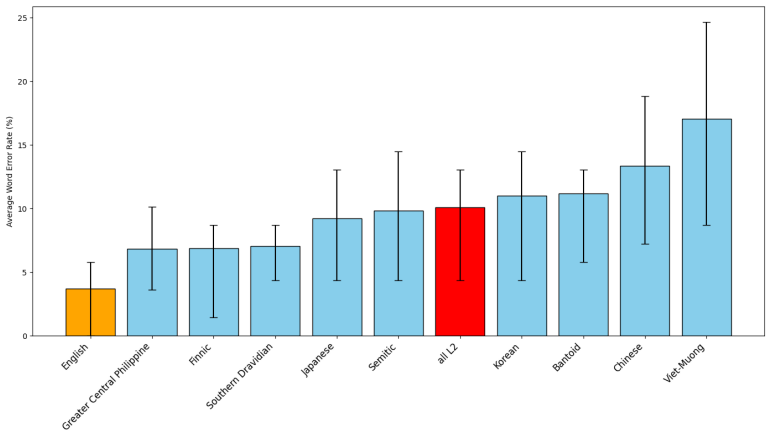



Figure 5.4: average WER for a variety of Genera with many datapoints in the corpus

All of the groupings in the above table were statistically significantly different from the L1 group. The selected groups tended to be the plurality genera from each of the language families in the above. For some of these if there was a member such as this in the linguistic family, it would have a WER that was slightly lower than that of the linguistic family on its own. This can be seen here with Greater Central Philippine language groups, which includes the most commonly spoken languages in the Philippines, Tagalog and Cebuano. While the Austronesian Group has a WER of 8.5%, the Greater Central Phillippine languages have a WER of 6.9%. There is a posibility that the greater visibility of these more common languages leads to situations where speakers of these languages are better represented in ASR models than their linguistic cousins who being more obscure have less recordings of themselves that ASR models can be trained on.

However, This is not always the case with plurality languages. For example The Viet-Muong, or Vietic Langage Genus, which includes Vietnamese, has a WER of 17.1% while the austroasiatic languages as a whole in this dataset have a WER of 16.5%, making it a worse performer than other languages in the Family such as Khmer in spite of much larger global representation. However, given the count of the Austroasiatic speakers is only 50, and Viet-muong make up about two thirds of that number, it is difficult to say if that is a true trend or limited to just this dataset.

One thing worth noting is the WER averages of L2 speakers of Japanese languages, Korean languages, and the Chinese languages. Noticeably, the Japanese Lanuages are the only one of the 3 to outpeform the L2 English speaker average, with a WER of 9.2% , whereas Korean trails Japan by almost 2 percentage points (11.0%) and Chinese has a much larger 13.4% WER. This higher performance on Japanese over other languages was also reported in Markl's analysis of bias in ASR systems, indicating that Whisper may still display the similar biases as older ASR systems. Given our understanding

of what makes a ASR system better at transcribing, we can make some guesses as to what may cause this discrepancy in WER. Firstly Japan and also Korea's closer national relations with Anglophone countries probably results in a much higher number of ASR resourses available from speakers of that language. This would of course include data of transcripts of L2 English speakers. Another Reason could be that due to larger cultural exchange between Anglophone countries and these countries, people in these countries are more likely to make Online videos in L2 English, leading to a more robust database for Whisper to train on, as compared to Chinese L2 English speakers, who are most likely less represented in ASR training sets.

#### 5.1.1.1 Bias present between first language English speakers

In this Experiment, we attempt to break the English language down into groups and analyse them further. Unlike the Language analysis, we could not use the information in the WALS to separate out subgroups, as it does not contain English dialect information. The SAA dataset does not contain self reported Accent information, so we made assumptions of accent based on place of origin, and split the English speakers into categories. Although this was not perfect, Over the course of dozens of speakers, this assumption would most likely hold true enough that outliers would only shift the values slightly.

At this stage an issue arose when analysing the English language data, in that the provenance of the SAA (from GMU) meant that of the 660 audio recordings of L1 language speakers, 433 of these recordings were recordings of Natural born residents of the United states of America. Without the anthropological expertise to be able to analyse the numerous dialects of American we are forced to grouped together most of the files into the 'American' category. The dataset is not suited to analysis of bias within American speakers, a task that is much better attended to by comparative corpora such as CORAAL.

As smaller linguistic origin points such as locations in the Caribbean were harder to put into a single group without manual labelling of all data files, we decided to split off all categories that contained at least 10 speakers in the corpus. This was the US, UK, Canada, Australia, Ireland and Jamaica. Additionally, as a majority of the remainder of nations represented English 1st language speakers from places that are traditionally less represented, New Zealand, as a country often considered a 'core' Anglosphere nation had a much lower WER than the rest of the group so was merged with Australia, to represent a 'antipodean' English group.

The remainder were placed in a miscellaneous group that included English 1st language speakers from a variety of backgrounds including South Africa, The Caribbean, Panama, South Asia amongst other locations. This group, while not unified would represent the WER of native English speakers from backgrounds not traditionally considered the 'Core' Anglosphere. All groups were statistically significantly different from the US Accent group except the Canadian and Australian sets

Analyzing the results, it is very easy to see that the North American Accents score the best with Whisper. The Canadian accent performs better than the US accents, but

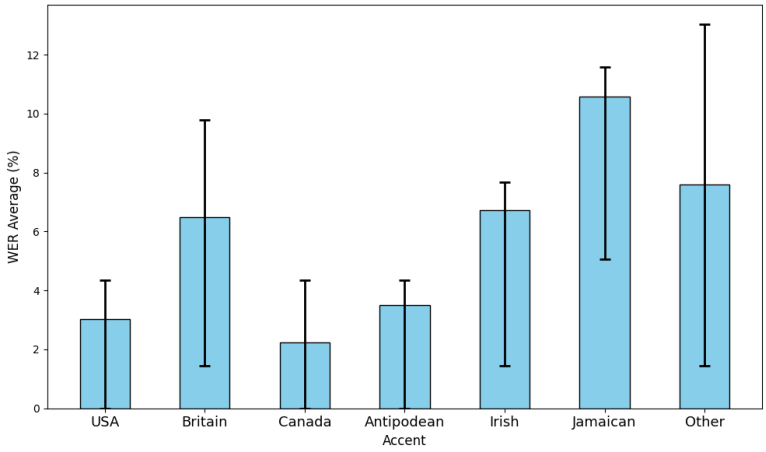



Figure 5.5: Wer Average for L1 English speakers, by country of birth

not significantly. These 2 groups have the lowest WER of any group isolated in this report. As the majority of English content on online video repositories is recorded and presented in American English, this is probably down to the great representation in the training data. Australian performs well, but not statistically significantly so. British English, while still outperforming the average L2 speaker, seems to perform very poorly compared to an expectation based on previous work in ASR bias. It performs equally as well as the Irish speakers, which opposes findings of Markl, from the background.

A look into the corpus gives us some theories. The dataset seems to contain very little in terms of speakers from the southeast. The corpus only contains 2 speakers from London, with greater variety in accents that a representative sample . This is probably very useful for the primary purpose of the SAA, to be a record of dialects and accents for linguists. Ultimately, this causes the value for the British speakers to be mainly useful for seeing performance by Whisper on less common English dialects.

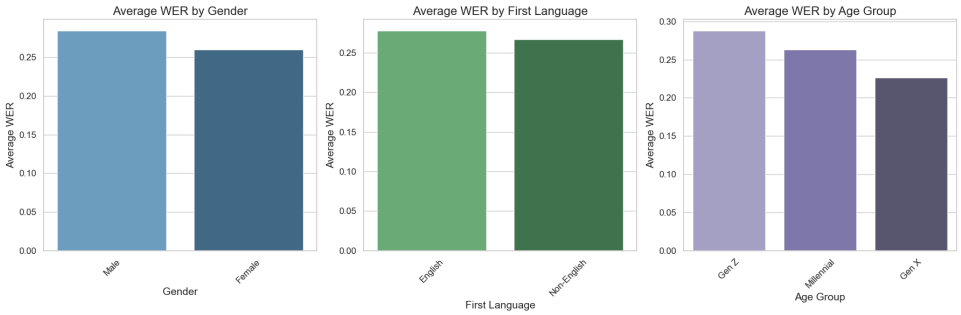
Irish also appears to be something that the Whisper model struggles with, with speakers from Ireland having a WER double that of an American speaker, But Jamaican accents in particular appear to give Whisper a lot of trouble, with the highest WER of all of the isolated groups in the English Dataset.

Notably, the quartile values for the Other L1 category are huge and widely varied. This indicates that this group fails to represent a group of speakers with a similar WER at all, but it is difficult to break this group down further as most of the countries where these English speakers reside only have 1 speaker each.

### 5.1.2 Analysis of Base transcripts on the EdAcc corpus

Compared to the SAA, the EdAcc is a corpus that represents a smaller number of speakers, But instead has much longer audio recordings, the advantages of which are explored in the dataset section. In this section we test the EdAcc database using a Whisper model of 'base' size, and look for points in the data where we are able to identify that a difference in the quality of transcription is evident, signaling a biased performance by Whisper.

The 3 graphs below are tables generated using the values in EdAcc's comprehensive speaker table to divide the corpus and identify using broad categories if there is an noticeable difference in the WER. Although the size of the database is much smaller,




there are still enough data points to be able to draw results. Firstly, in this database, The WER value difference for male speakers and female speakers was too small to indicate difference, ($p > 0.05$), and the same can be said for the transcription rate of speakers who specified that they were first language speakers of English against those who did not list English as one of their first language, as interesting as the fact that 1st language English speakers seemed to have a worse WER is. The difference in the WER between GenX and GenZ was significant ($p < 0.05$), and there appears to be an observable decrease in the quality of the transcription as the speaker gets younger, which agrees with previous analysis of WER by age from the background. This may be due to older manners of speech being mostly identified as more 'standard' with more data, while younger speakers may speak actively developing dialects with less data.

It is important in an analysis of bias to try to identify places in the Whisper model where bias is not detectable, although these results are less interesting than the discovery of bias. However, this does not mean that there is no bias, but that we have not detected that bias. It is important to understand the greatly decreased number of results for every category due to the corpus size. The fewer individuals there are in a group the more likely the average value is to be swayed by a few unrepresentative members, meaning that the observed difference in WER between categories may be due to a third variable.

Using the self reported Accent information for each of the speakers and grouping together comparable or related langauges below we plot the WER of different Accent groupings across the EdAcc corpus. Issues faced in the use of Accents are discussed

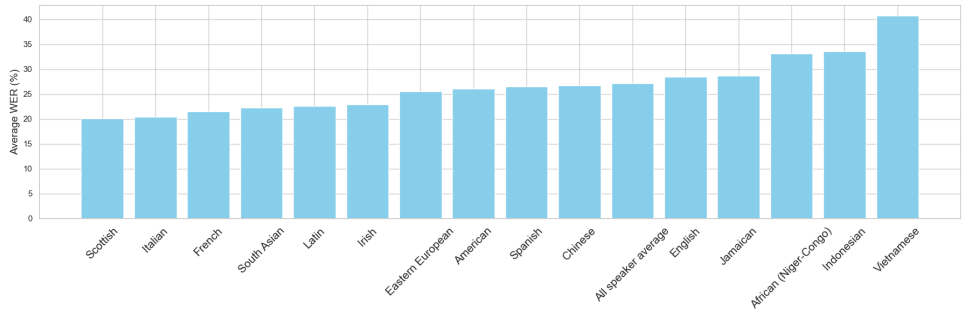



Figure 5.6: WER Values for accents in the EdAcc corpus

in the methodology section. Here they manifest themselves in unusual scoring for certain groups in the Corpus. For example, English and American score unusually high in comparison to what one would expect. This can be explained by looking at the information of users that identify their accents as English or American. Often, the descriptor 'American' is used by European speakers of English, potentially to use a more familiar label when describing their accent. With English the low result appears to be due in part thanks to the EdAcc's curated strength in identiying bias in the abilities of ASR systems. Of the Speakers who identified themselves as having an accent that could be classified as English, Many of themselves were of groups who were most at risk of bias in ASR based on historical ASR testing. Of the 5 worst scoring values in this subset, only 1 speaker was of non-white ethnicity, and many identified as speaking regional accents such as "South London". Additionally, the ethnically white speaker with a high WER had a history of eardrum surgeries and hearing issues. Due to these values, The English and American values in the above table do not act as benchmark values that represent Whisper on 'standard dialects'.

With the exception of the English and American values in the above table, The rest of the results in the table strongly corroborate the results as seen in the SAA, with Vietnamese being the language that struggles the most with recognition by Whisper in both datasets, as well as the relative WER distances from language groups being mainly preserved, with lower WERs for South Asian and Romance Language speakers and higher ones seen for Chinese, and African (Niger-Congo) accents. Additionally, the relative scoring of Romance languages also appears to be preserved in these results.

We realised that a further analysis into the set of English speakers was required. In the table below, we have separated out all speakers in the corpus who list English amongst their Primary languages, and then calculated the WER for each identified Ethnic background.

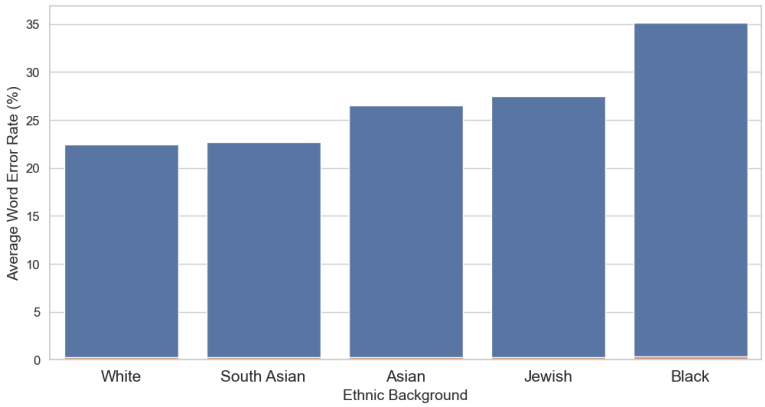



Figure 5.7: WER averages by Ethnic Background for reported first language English speaker in the EdAcc

The results show a clear bias in the performance of Whisper on individuals of certain ethnic backgrounds over others, with a performance difference of 12.5% between White and Black speakers of English. Of the Ethnicities in the table, The difference between White and Black speakers is the only one where the difference is statistically significant.

($p < 0.05$). This could be caused by a variety of factors such as greater prevalence of dialects that Whisper struggles with within these populations, which ultimately result in unequal performance received by ethnically Black individuals on average.

## 5.2 Analysis of difference in transcription quality between different versions of Whisper

As stated in the methodology, Whisper actually can be downloaded in a variety of sizes, smaller ones conferring faster transcription times and smaller load, and larger models promising more accurate transcriptions.

### 5.2.1 Testing on the SAA

In this section we calculate the WER for each file in the SAA as they move from smaller models to faster models, with an interest in whether this rate of improvement is consistent between languages, between the different model sizes

The table below gives the assumed relative speed of the Whisper model against the time taken for the Whisper model to transcribe all 3032 files in the SAA, totalling 22 hours of audio.

| Model Size | Time Taken (Hours), SAA | Expected Relative Speed | RAM (GB) |
|---|---|---|---|
| Tiny | 0:37:35 | 16x | ∼1 |
| Base | 1:26:40 | 8x | ∼1 |
| Small | 2:34:40 | 3x | ∼2 |
| Medium | 7:48:02 | 1x | ∼5 |

Table 5.3: Model performance and resource usage.

Note that the values of time here are not a true representation of Whisper at optimal performance, as these times were not taken under controlled circumstances. The data demonstrates a clear trade-off that is made when using different sized Whisper models. Given the increasing load on a computer and time required to transcribe audio, for most purposes, such as any applications of Whisper that require faster times, or more portable forms of Whisper that may be using less powerful hardware, the smaller forms of Whisper will be used. Therefore, there is a strong value in an analysis of bias that may be found in the smaller models.

Below is a table illustrating the curve of the average WER for the entire corpus as the model improves, alongside another table with every language plotted as a line, Showing the range of different rates of improvement that can be seen depending on the first language of the L2 speaker.

The first image we include to give a baseline understanding of the improvement in transcription quality, as the trends shown on this chart roughly hold for all data in the dissertation. Notably, you can see that there is a decent improvement from the tiny to base model, a sizeable improvement from base to small model and a much smaller

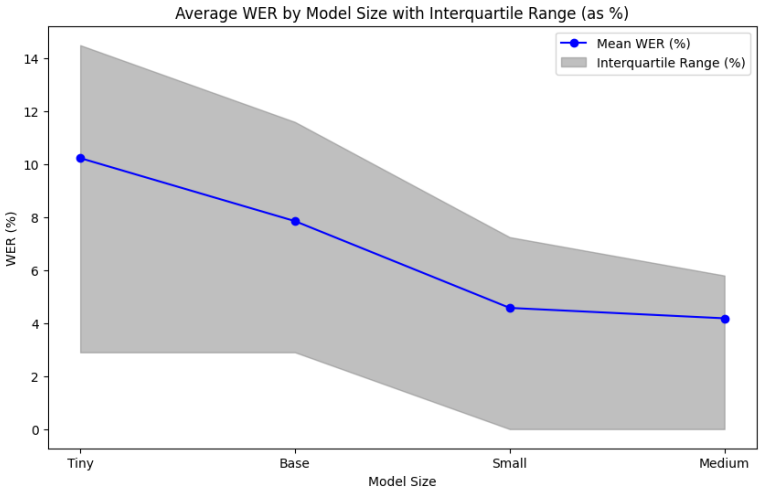



Figure 5.8: Plot of the mean WER of the SAA as Model Size increases, with interquartile range shaded

return on quality on the small to medium one. The WER sees a 23.3% improvement from tiny, a 41.7% improvement from using the small model and a 8.6% improvement from small to medium. The grey space represents where the line could be for most of the languages within the corpus. One can make out the general trend of improvement, and notice that for the most part, many data groups do follow this pattern of improvement of values. The Bias in Whisper in decrease as higher quality transcriptions are produced by larger Whisper models. However, as you can see from the height of the grey section, not all languages follow the line of improvement, and can improve at different rates, or not see the improvements in WER that other languages can see from larger Whisper model sizes.

Therefore, in the image below, we have selected the genera from the previous experiments and reused them to track the performance of Whisper on large groups of similar first language speakers

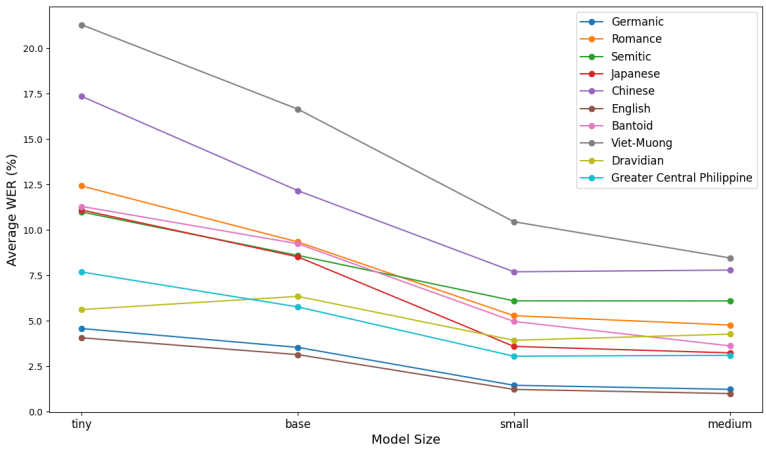



Figure 5.9: Plot of the average WER of different Linguistic Genera as the Model size increases

English is unsurprisingly the best performing L1 group, consistently getting a WER lower than all other linguistic backgrounds. Germanic L1 speakers performs close to English, hovering roughly 0.2% above the English WER in all models. Beyond this, the bias in the Whisper model that was previously analysed becomes apparent, and larger gaps in the word error rate between languages open up. although the positions of each of the languages relative to each other is roughly the same as in the previous experiment, there are a variety of deviations from the line of improvement that are worth noting. Firstly, Dravidian L1 speakers seems to have a WER that does not correlate to the model size. Having analysed the values in the data frames, this is for no reason that we can ascertain, although a completely unusual line of transcription for just speakers of one language family would be a bias if this were a replicable result. However Dravidian is not the only language group that gets worse as a model improves. The Chinese and Philippine language groups also have a worse WER for the Medium model than they do for the small, although by a much smaller margin. Additionally, Japanese seems to improve at a much greater rate than other similar languages, but only between the base and small results. The Below Table supplements the data in figure 5.8 by displaying the ratios of the word error rates of each of the Language Groups between each of the model size.

Looking at the information in table A.1 in the Appendix which shows the ratio of the change in WER between models against the WER on the smaller model for each of the L2 groups, Not only does English already have the lowest word error rate of all language groups, but it additionally benefits the most from improved model sizes, with its improvement in WER from tiny to medium being the largest of all of the language groups shown. Additionally, the numbers in this chart show that at each movement to a larger model, the improvements between languages are very varied. Another angle from which to view this is from the perspective of a user of a larger model moving to a smaller model. Chinese has the largest ratio of the language genera when it moves from a small model to a base model, as does the Bantoid genus in the move from medium to small. A user from one of these L1 communities will be more affected by the move to a smaller model than other languages

There is bias in the Whisper model that is introduced by the selection of a model size: The increasing of a model size in Whisper does not necessarily impart a similar improvement in transcription for speakers of some languages. This can affect languages in many ways that make using a larger or a smaller model acts that may impart harm to subsets of the userbase.

### 5.2.2 Testing on the EdAcc

In this section of the paper we complete an analysis of the performance of different model sizes of Whisper on different accent groups within the EdAcc corpus. From the above testing on the SAA, we determined that the improvement seen from the use of the medium model was relatively unimpressive as compared to the improvement gain seen using the small model. As this corpus is longer in terms of raw length of audio, considering time for transcription it was decided that bias would be detected using just 3 model sizes as the improvement in WER from using medium would be marginal. Below

Is a table with accent groupings found in the EdAcc corpus, chosen based how much they improve between the base and the small model in order to illustrate the different rate of improvement between models for different languages.

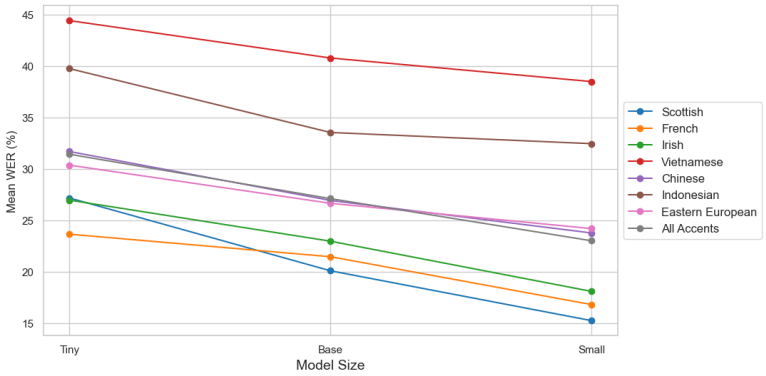



Figure 5.10: Chart plotting the WER of different accents in the EdAcc corpora as the model size increases

Looking at the above set of lines, , especially between the base and small model, one can see that the lines drawn across from the WER values are at different angles, indicating that some accents are seeing greater improvements at larger models, while other languages are not as improved. This is most easily seen when compared to the angle of the All accents line. Languages with good performances in smaller models such as Irish, Scottish accent speakers see a greater degree of improvement than speakers of Vietnamese, Chinese, Indonesian and Eastern European speakers. Another point in the data worth considering is the loss of WER for Scottish accents between the base and tiny model. As compared to the other languages, Scottish seems to worsen the most, meaning that the adoption of a lighter model would disproportionately affect speakers with Scottish accents, demonstrating bias.

## 5.3 Tests on the Whisper model's language detection

### 5.3.1 Testing on the SAA

Thus far in our experiments we have given Whisper the information of the spoken language of the input, which has always been English. As explained in the methodology, Whisper has the ability to work out the language of the incoming audio. It does this by assigning a score of 0 to 1 to each language, where all of the scores sum to 1, representing its confidence in the file being in that language. In this section, This ability of the Whisper model is tested, noting points of failure and if any specific languages or other are more likely to fail to be identified correctly.

Firstly, of the 3029 files in the SAA dataset, 211 files failed to be correctly labeled as English, representing a failure rate in this corpus of 7.0% to correctly identify the language being spoken due to the accent of the speaker. None of these files were L1 files.

#### 5.3.1.1 Decrease in quality of transcription due to failed transcriptions

Below is a chart that shows the difference of the average WER of these 211 files when they are correctly identified as English, and when they are identified as a foreign language.

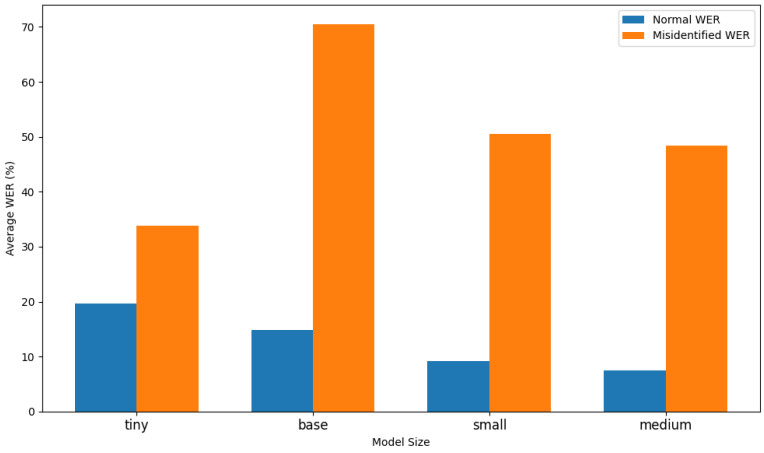



Figure 5.11: Chart that shows the degradation of the WER on misidentified files in the SAA at different model sizes

Whisper will still sometimes identify English words spoken by the individual correctly, meaning that the WER does not become 100% when Whisper begins to operate on foreign languages, but will mainly identify words from the detected language within the spoken words instead. As can be seen from the above results, with the exception of the tiny Whisper model, the average WER balloons to 4 times that of the correctly identified WER, causing Whisper to become all but unusable to these individuals.

A look at the confidence scores that were given to each file also reveals that for 10 of the files, Whisper had a greater confidence interval between primary guess and English of 0.9, and 29 with an interval of $>$0.75, Most of the files, however fall below 0.35, indicating relatively little confidence in the identification of the file.

Below is a table with the languages that shows the 6 languages with the highest false language rate, out of the languages with more than 10 speakers in the corpus

| | Hausa | Hungarian | Tagalog | Bengali | Japanese | Macedonian |
|---|---|---|---|---|---|---|
| Error Rate | 0.30 | 0.23 | 0.22 | 0.21 | 0.20 | 0.19 |
| File Count | 10 | 13 | 27 | 24 | 45 | 27 |

Table 5.4: Table showing 6 most frequently misidentified L2 Speakers

This is actually surprising, because Whisper is showing Bias but to different languages and groups than it has previously done so. Note that Japanese was an unusually well performing language in the previous tests, but here is the 5th most commonly misidentified L2 English speaker. This does however show that certain L2 speaker

groups are more likely to get misidentified with Whisper due to the classifier having low familiarity of their accent of English, which causes these groups to disproportionately fail to be transcribed

### 5.3.2 Testing on the EdAcc

This section repeats the process completed above on the EdAcc to explore the impact of mistakes made by a language identifier. Although the files in the EdAcc are longer than those of the SAA, The language identifier is unable to use any of this additional information, as it is designed to deduce the language spoken using the first thirty seconds of audio only.

Due to the much lower speaker count in the EdAcc, only 4 files were found to be misidentified as audio of individuals speaking a language other than English, which represents a failure rate of 3.2%. The clips in question were in no particular order, a male black L2 French speaker, a female white L2 Catalonian Speaker, A male black L1 English and Jamaican Creole speaker, and a male black L2 Igbo speaker. The result of being misidentified by Whisper is that Whisper was completely unable to transcribe these speakers, with the WER average between these speakers reaching 90%. Although this is a relatively small sample, this is evidence of the Bias of the Whisper model against these individuals and others who share a similar linguistic background that would cause a difference in experience for different users of the same service.

## 5.4 Whisper on audio samples with noise

Whisper is as previously explored by OpenAI, much improved compared to contemporary models on transcribing audio on a variety of noise levels. This allows it to still transcribe when the speaker is in less than ideal environments.

### 5.4.1 Testing on the SAA

While transcription being possible at these higher SNR's is definitely better than the alternative of no transcriptions at all, the Whisper model may demonstrate bias towards certain L2 English speakers by affecting them differently as the noise level increases.

The above Figure shows what the general curve of the decrease in WER looks like. The further the curve to the left, the worse the language performs in noisy conditions. Table A.2 in the appendix is a table that represents the increase of the WER between different SNR values in figure 5.13. In the first column looking at the increase in WER from 30-20 SNR, Japanese is the language that is affected the greatest. At an SNR of 20, Dravidian has the largest increase in SNR. Here, the difference in the values for the WER increases really differentiate themselves. The English increase in WER is substantially lower than that of the other languages, meaning that the transcriptions are degrading due to the noise at a faster rate than for the L1 English speakers. In the final column of 10 to 0, the largest jump is English, due to the fact that it was the only L1 group remaining that still had far to go before total failure to transcribe.

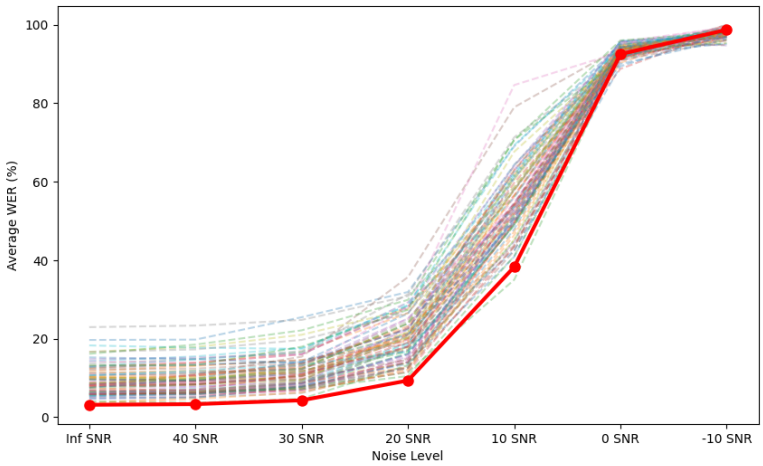



Figure 5.12: Illustrative chart showing all languages WER at different SNR values. English In Red to highlight L1 performance against L2

All of the features described above in the table can be seen in the graph below, as German and English seem to perform better in a noisy environment than other language groups which all seem to follow a worse gradient across from 30 SNR to 0 SNR. This represents a bias in Whisper of higher performance for L1 and German L2 speakers. This could occur because there is not enough noisy data in the training data for speakers of other linguistic backgrounds, reducing Whispers performance for these other linguistic backgrounds.

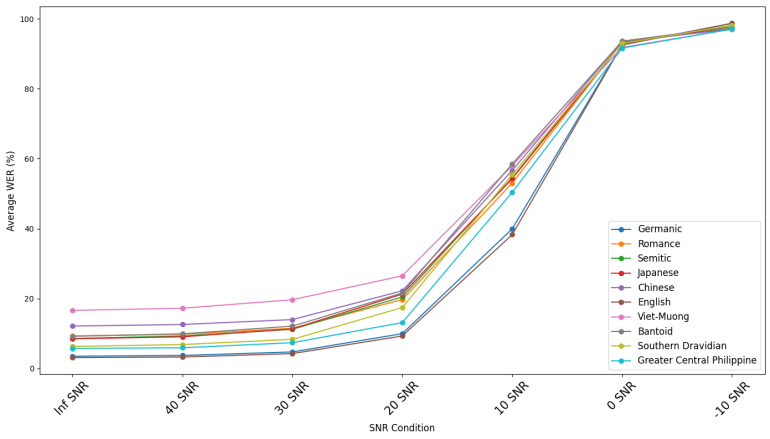



Figure 5.13: Figure showing the degradation of the Whisper transcription for different linguistic genera as the SNR is increased

### 5.4.2 Testing on the EdAcc

This section focuses on the testing of the accent groups found in the EdAcc using the same test methodologies as the SAA in the above section.

As mentioned in the methodology section, the WER calculations used for the EdAcc and SAA are different, with measurements of WER on the SAA not including insertions,

while measurements on the EdAcc do. The result of this is that in the plot below, should the number of insertions be high enough, the number of errors in the Whisper transcription will be higher than the number of words in the gold standard transcription, resulting in a WER greater than 100

The table below plots the degradation of the quality of transcriptions in Whisper for different self reported accent groups across the EcAcc corpus as the noise level is increased. Due to crowding the table, some accent groups are not on the plot, as they follow similar trajectories to other accent groups.

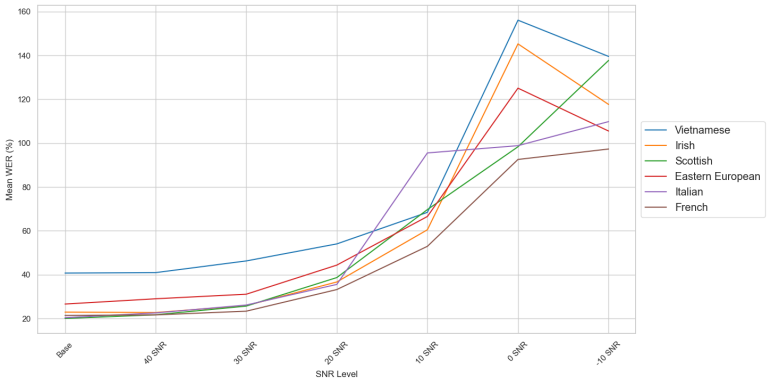



Figure 5.14: Figure showing the increase of WER as noise levels are increased for different accent groups in the EdAcc corpus

The accent groups South Asian and Chinese followed line trajectories similar to that of Vietnamese, while Irish, Latin American and Indonesian followed line trajectories closer to that of Scottish. These groupings seem odd. The Variety of trajectories that the WER values follow shows that there is some bias demonstrated by Whisper on its ability to transcribe at Higher SNR values, but these values do not line up with biases noted during other analysis throughout our testing. Possibly a relatively high WER value does not preclude the training set from containing noisy transcribed data for some of these L2 speaker groups.

From the languages that are on the chart, we can see that some language groups are more prone to hallucinations by Whisper at high enough noise levels compared to others, which could cause speakers of certain accents in noisy environments to be transcribed to say completely different things than what they said. The presence of hallucinations in the WER values is seemingly unrelated to the rate of degradation of transcription up until about 10 SNR, as is seen from the French and Irish lines. Italian seems to have induced hallucinations sooner than other language groups, showing that this fault in Whisper can occur at lower noise levels for speakers with certain accents. Hallucinations can be worse than simply being unable to be transcribed depending on the content of the hallucination, which could be a path into further research in the topic of ASR.

# Chapter 6

# Conclusions

This project set out to Explore the bias in very large automatic speech recognition systems. By using WER as a metric and plotting different groups by linguistic background, we were able to demonstrate multiple cases where Whisper demonstrated differential performance on certain groups, indicating bias on Whisper's part.

To accomplish, we started by gathering and pre-processing our datasets. This was an involved process that involved scraping a website to retrieve files as well as metadata for the speakers for 3032 different files for one of the corpora, and writing code that enabled the splitting of dyadic audio into individual halves for another.

Following on from this, we researched and studied linguistic metrics and linguistic databases that could be used to classify the speakers into larger groups to enable the identification of trends in data with greater validity

In our methodology section we laid out experimental details and clarifications of concepts required for completing the tests.

This lead to the actual experiments conducted in this paper, Which included a study of bias on 2 datasets on 4 different sizes of the Whisper models under multiple circumstances as well as testing at 6 different SNR levels. This comprehensive evaluation resulted in the generation of 12 sets of transcriptions for the EdAcc corpus and 14 for the SAA corpus.

We completed in depth tests that analysed the performance of the 'base' Whisper model, finding that there was great variation in how well Whisper was able to transcribe based on factors such as the primary language of the speaker, identifying groups of L2 speakers that Whisper found the most difficulty with. We also found Groups within L1 Speakers that received diminished WER values from Whisper, indicating bias.

We tested Whisper models to identify the extent to which Whisper is more or less biased at different model sizes, as well as noting the difference of the improvement between speakers when moving between models, Which would indicate biased performance for different speaker groups.

We also tested the language detection functionality of Whisper and found that it was not

a reliable means of determining the spoken language of a clip, a flaw that was causing Whisper to greatly mis-transcribe speakers. We made attempts to identify if there were trends between misidentified files.

Finally we also analysed the degradation of transcriptions as noise levels were increased in clips within the corpora, and analysed which languages were able to retain high WER as others soon became impossible to transcribe under a given noise environment, finding That L2 Speakers of certain languages would show worse WER values at higher SNR values than other L2 and L1 speakers, and noting that hallucination of added words would occur for certain speakers given a high enough noise level.

Through rigorous testing and analysis this dissertation has achieved the objectives set forth in the introduction. We have created compelling evidence of bias in the Whisper models across multiple scenarios, with specific examples of failure documented, such as the overall poorer WER for L2 speaker against L1 speakers

This work substantiates the need for further investigation into bias in Whisper, with many potential paths of exploration. This includes such as phone or feature level analysis of transcription errors by Whisper as inspired by related literature, or potentially an analysis into the hallucinations Whisper seemed to generate at sufficiently high noise levels, and whether this changes by features of the speaker in ways that demonstrate bias.

# Bibliography


[1] P. Bell, O. Klejch, A. Carmantini, N. Markl, N. Bogoychev, and R. Sanabria Teixidor. The edinburgh international accents of english corpus, 2023. URL https://doi.org/10.7488/ds/3832.

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

[3] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix Smits. JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/254$_pdf.pdf$.

[4] Matthew S. Dryer and Martin Haspelmath. Wals online, 2013. URL https://wals.info. Available online at https://wals.info. Accessed on 2024-04-07.

[5] Siyuan Feng, Olya Kudina, Bence Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition, 03 2021.

[6] Esther Grabe and Brechtje Post. Intonational variation in the british isles. 03 2002.

[7] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Alexey Karpov, Oliver Jokisch, and Rodmonga Potapova, editors, *Speech and Computer*, pages 198–208, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99579-3.

[8] Eric W. Holman, Author's First Name Adelaar, Author's First Name Blust, and Author's First Name Campbell. Automated dating of the world's language families

based on lexical similarity. *Current Anthropology*, 52(6):841–875, 2011. doi: Insert DOI if available. URL Insert URL if DOI not available.

[9] International Organization for Standardization. Iso 639-1 and 639-2 :2023 codes for the representation of names of languages. ISO Standard, 2023. URL https://www.iso.org/standard/74575.html. Accessed: date-of-access.

[10] B. Juang and Lawrence Rabiner. Automatic speech recognition - a brief history of the technology development. *Journal Name Here*, Volume Number Here:Page Range Here, 2005.

[11] Tyler Kendall and Charlie Farrington. The corpus of regional african american language. https://doi.org/10.7264/1ad5-6t35, 2023. Accessed: date-of-access.

[12] Jason Kincaid. A brief history of asr. Medium.com, July 2018. URL https://medium.com/descript/a-brief-history-of-asr-automatic-speech-recogni [Online; posted 27-August-2012].

[13] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

[14] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947. doi: 10.1214/aoms/1177730491.

[15] N. Markl. Language variation and algorithmic bias: Understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, pages 521–534, Seoul, Korea, Republic of, 2022. ACM Association for Computing Machinery. doi: 10.1145/3531146.3533117. URL https://doi.org/10.1145/3531146.3533117. 5th Annual ACM Conference on Fairness, Accountability, and Transparency.

[16] Joshua Martin and Kevin Tang. Understanding racial disparities in automatic speech recognition: The case of habitual "be". 10 2020. doi: 10.21437/Interspeech.2020-2893.

[17] National Institute of Standards and Technology (NIST). Sctk: Speech recognition scoring toolkit. NIST, 2023. URL https://www.nist.gov/itl/iad/mig/tools.

[18] Nelleke Oostdijk. The spoken dutch corpus: Overview and first evaluation. *Proceedings of LREC-2000, Athens*, 2, 01 2000.

[19] OpenAI. Announcing the large-v2 model. GitHub repository discussion, 2023. URL https://github.com/openai/whisper/discussions/661. Accessed: date-of-access.

[20] OpenAI. Announcing the large-v3 model. GitHub repository discussion, 2023. URL https://github.com/openai/whisper/discussions/1762. Accessed: date-of-access.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel. The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 01 2011.

[23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[24] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[25] Miguel Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jette. Earnings-21: A practical benchmark for asr in the wild, 04 2021.

[26] Miguel Rio, Hendoro Peter, Quinten McNamara, Corey Miller, and Shipra Chandra. Earnings-22: A practical benchmark for accents in the wild, 03 2022.

[27] David Sankoff. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparisons*. Addison-Wesley Publishing Company, Inc., 1 edition, 1983.

[28] Maurizio Serva and Filippo Petroni. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81, 08 2007. doi: 10.1209/0295-5075/81/68005.

[29] S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, dec 1965. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

[30] Stanford Linguistics. Voices of california. Stanford University, 2023. URL http://web.stanford.edu/dept/linguistics/VoCal/. Accessed: 07/02/2024.

[31] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[32] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. pages 53–59, 01 2017. doi: 10.18653/v1/W17-1606.

[33] Nik Vaessen. jiwer: a python package for evaluating automatic speech recognition systems. URL https://github.com/jitsi/jiwer.

[34] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[35] Steven Weinberger. Speech accent archive. George Mason University, 2015. URL `http://accent.gmu.edu`. Retrieved from George Mason University website: `http://accent.gmu.edu`.

[36] Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The asjp database. Online Database, 2022. Accessed: date-of-access.

# Appendix A

# First appendix

Below is the Elicitation Paragraph used in the SAA.

> Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

| a:b<br>Language Group | tiny:base | base:small | small:medium | tiny:medium |
|---|---|---|---|---|
| English | 0.228 | **0.613** | 0.189 | **0.758** |
| Germanic | 0.227 | 0.592 | 0.156 | 0.734 |
| Japanese | 0.233 | 0.580 | 0.099 | 0.709 |
| Bantoid | 0.182 | 0.464 | **0.271** | 0.680 |
| Romance | 0.248 | 0.435 | 0.098 | 0.617 |
| Viet-Muong | 0.218 | 0.373 | 0.191 | 0.603 |
| G.C.Philippine | 0.250 | 0.472 | -0.015 | 0.598 |
| Chinese | **0.299** | 0.368 | -0.012 | 0.551 |
| Semitic | 0.218 | 0.291 | 0.001 | 0.446 |
| Dravidian | -0.129 | 0.382 | -0.086 | 0.241 |

Table A.1: Table showing the ratio of the change in WER from a to b, and the WER when at model a for a variety of language groups

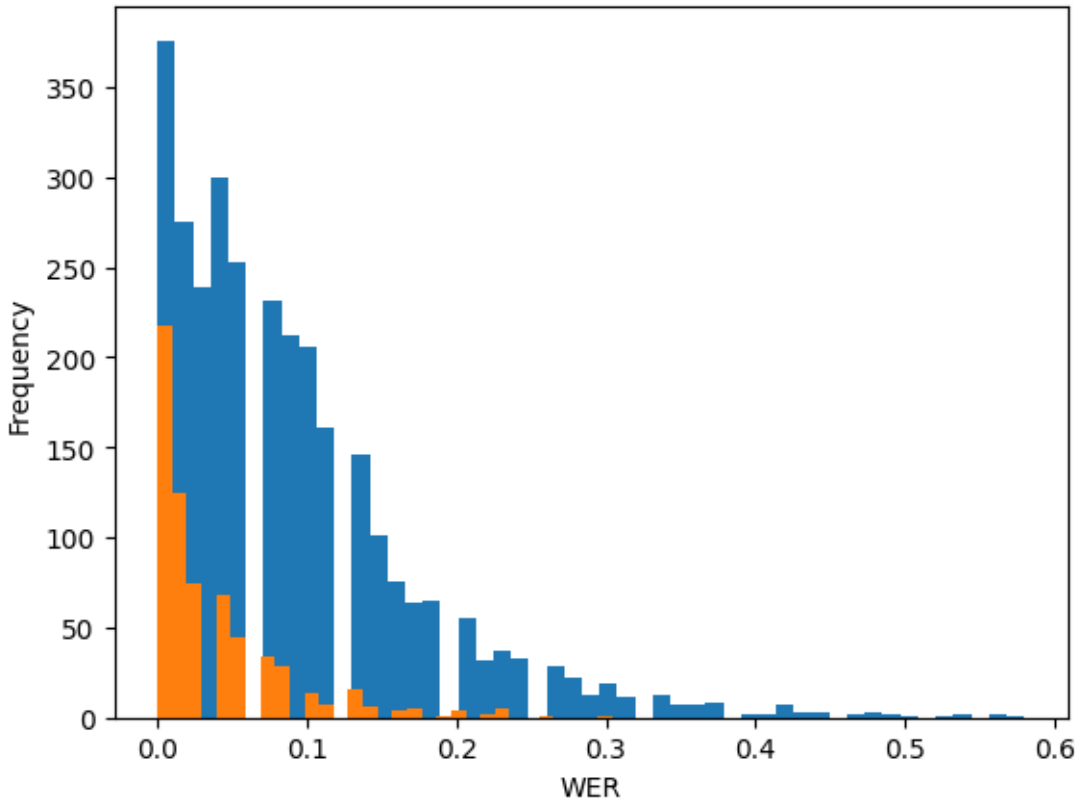



Figure A.1: Histogram of WER values in the SAA corpus, with English files in orange, and all others in blue

| Language Group | 30 to 20 | 20 to 10 | 10 to 0 |
|---|---|---|---|
| Germanic | 0.052 | 0.300 | 0.527 |
| Romance | 0.082 | 0.333 | 0.402 |
| Semitic | 0.092 | 0.341 | 0.388 |
| Japanese | **0.100** | 0.330 | 0.388 |
| Chinese | 0.082 | 0.344 | 0.370 |
| English | 0.050 | 0.290 | **0.542** |
| Bantoid | 0.095 | 0.369 | 0.351 |
| Viet-Muong | 0.068 | 0.315 | 0.338 |
| Dravidian | 0.091 | **0.380** | 0.376 |
| Greater Central Philippine | 0.058 | 0.373 | 0.412 |

Table A.2: Table showing the ratio of the change from SNR a to SNR b against the WER of SNR a

| Language | Count |
|---|---|
| english | 660 |
| spanish | 243 |
| arabic | 201 |
| mandarin | 157 |
| korean | 99 |
| french | 85 |
| russian | 82 |
| portuguese | 69 |
| dutch | 54 |
| turkish | 45 |
| japanese | 45 |
| german | 45 |
| italian | 40 |
| vietnamese | 40 |
| farsi | 39 |
| polish | 39 |
| cantonese | 36 |
| hindi | 34 |
| urdu | 29 |
| macedonian | 27 |
| tagalog | 27 |
| amharic | 27 |
| romanian | 24 |
| bengali | 24 |
| swedish | 23 |
| thai | 23 |
| bulgarian | 20 |
| serbian | 19 |
| nepali | 18 |
| gujarati | 17 |
| estonian | 17 |
| greek | 17 |
| telugu | 15 |
| finnish | 15 |
| ukrainian | 14 |
| tajiki | 14 |
| indonesian | 14 |
| albanian | 14 |
| tamil | 13 |
| hungarian | 13 |
| bosnian | 12 |
| punjabi | 12 |
| miskito | 11 |
| kiswahili | 11 |
| marathi | 11 |
| dari | 11 |
| taiwanese | 11 |
| hebrew | 11 |
| mongolian | 10 |
| pashto | 10 |

Table A.3: 50 largest languages in the SNR corpus and their counts